numero 153
maggio 1981
anno XIV
volume XXVI

LE SCIENZE
edizione italiana di SCIENTIFIC AMERICAN

Le eruzioni del Mount St. Helens

L'attuale ciclo eruttivo del vulcano si ascrive a un tipo di attività che si manifesta da 4500 anni. In effetti, le violente eruzioni dello scorso anno poterono essere previste in loco dai vulcanologi

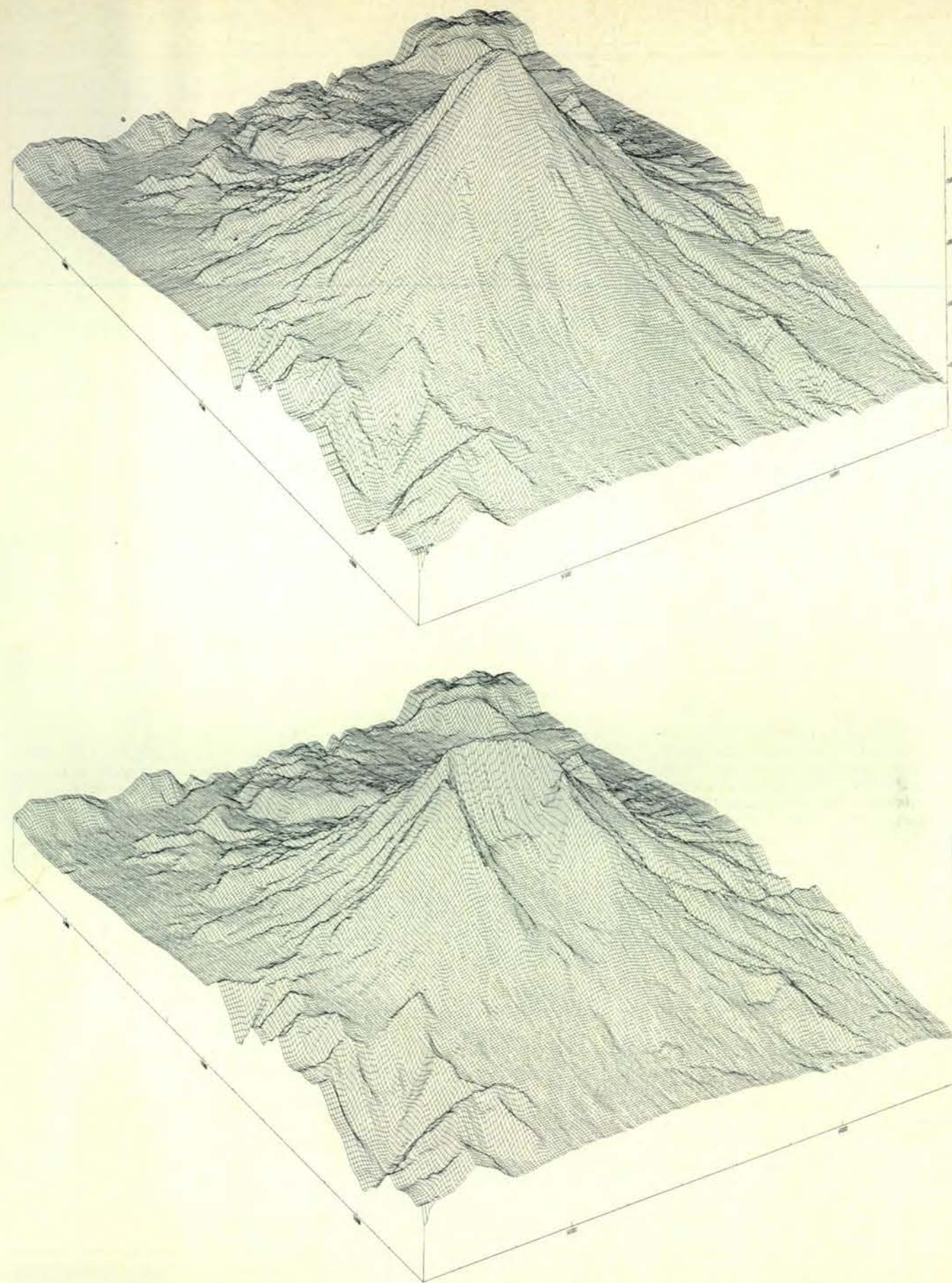
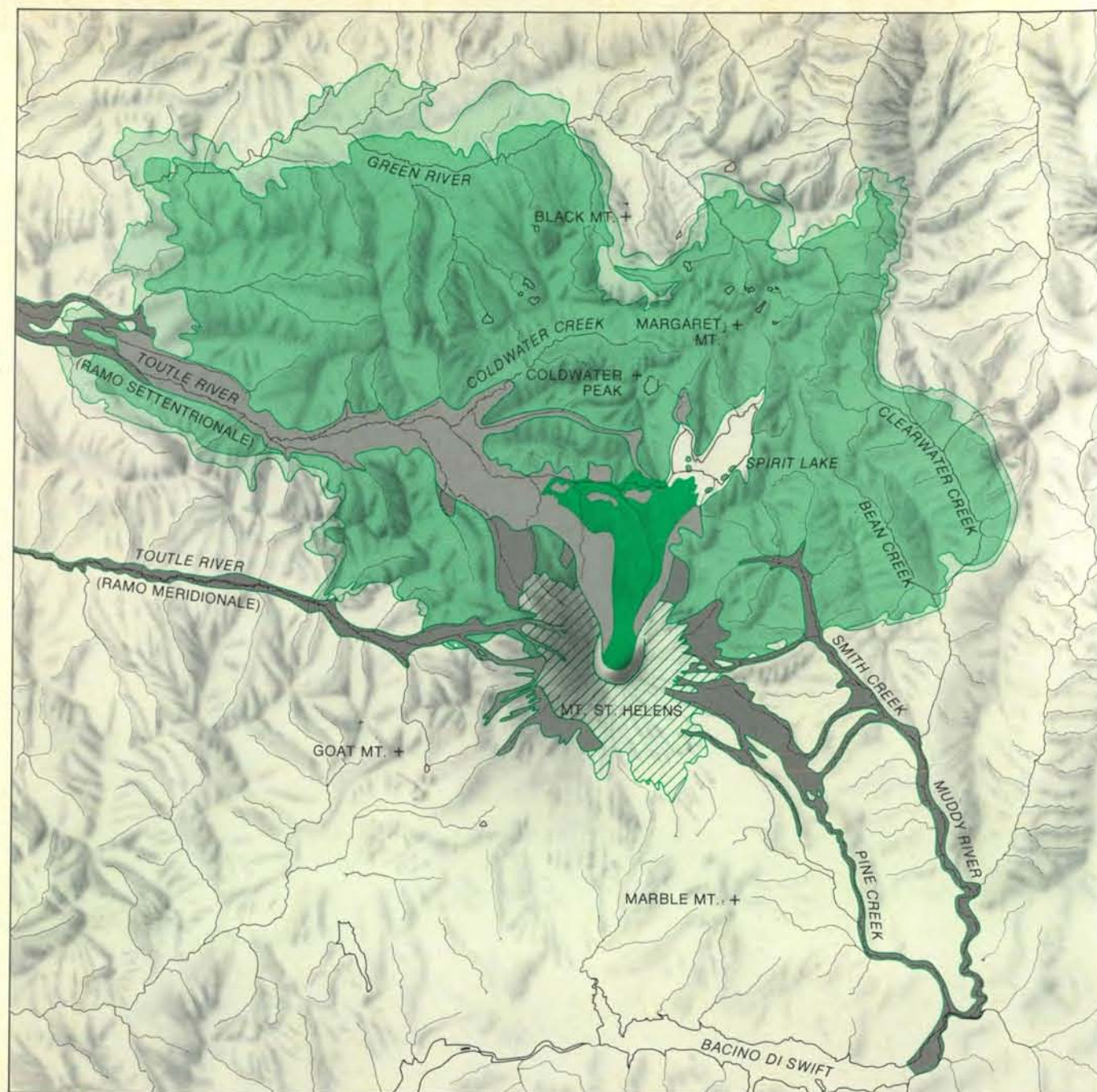
di Robert Decker e Barbara Decker

La violenta eruzione del Mount St. Helens del 18 maggio dello scorso anno è stata una delle più attentamente controllate, più intensamente fotografate, più ampiamente documentate della storia. Per il territorio degli USA tale eruzione è anche stata la prima nel complesso dei 48 stati contigui, dopo le eruzioni - molto meno violente - del Lassen Peak fra il 1914 e il 1917. L'eruzione di Mount St. Helens ha spostato 2700 milioni di metri cubi di roccia vulcanica (compresi 500 milioni di metri cubi di nuovo magma), devastando un'area superiore a 500 chilometri quadrati e dando origine a una delle più grandi valanghe della storia. Negli ultimi cento anni l'attività esplosiva del Mount St. Helens era stata nettamente superata come intensità soltanto dalle eruzioni del Santa Maria in Guatemala (1902), del Krakatoa in Indonesia (1883) e del Katmai in Alaska (1912), i quali espulsero rispettivamente circa 5000, 6000 e 12 000 milioni di metri cubi di magma (essendo questi i volumi relativi alla densità della roccia solida).

Anche il Katmai era stato rimpicciolito da precedenti eruzioni, come quelle che nel Giappone, nella Nuova Zelanda, nell'America Centrale, negli Stati Uniti occidentali e in molte altre regioni vulcaniche del mondo, avevano sepolto migliaia di chilometri quadrati con una coltre di cenere e di roccia di spessore variabile da dieci a cento metri. Il volume di materiale espulso nel corso di queste grandiose eruzioni variò da 100 miliardi di metri cubi a più 1000 miliardi. Il vulcanismo preistorico superò forse ogni nostra immaginazione possibile o l'esperienza umana è dotata di una visione troppo ristretta perché possa essere tentato un giudizio? Gran parte dei geologi propende per la seconda supposizione. Probabilmente le eruzioni degli ultimi cento anni sono soltanto piccoli campioni dell'energia vulcanica che la Terra ancora possiede.

Durante i mesi dell'eruzione, l'energia liberata dal Mount St. Helens è stata spesso descritta in termini che sono propri dell'energia relativa alle esplosioni nucleari. Tale confronto è utile, ma piuttosto fuorviante. Non solo è del tutto diversa la sorgente dell'energia, ma diversa è anche la velocità o la potenza con cui essa viene liberata. L'energia termica e meccanica liberata il 18 maggio 1980 dal Mount St. Helens è stata di circa $1,7 \times 10^{18}$ joule. Considerando che un'esplosione nucleare dell'ordine di un megaton libera circa $4,2 \times 10^{15}$ joule, ne consegue che l'eruzione del Mount St. Helens è stata equivalente a un'esplosione nucleare da 400 megaton, ossia quasi otto volte più potente di quella finora ottenuta da un ordigno nucleare. Il confronto è fuorviante perché in sostanza l'intera energia di un'esplosione nucleare si trasforma immediatamente in energia termica e meccanica, così che la potenza pressoché istantanea dell'esplosione nucleare espressa in watt è essenzialmente la stessa dell'energia espressa in joule. (Un watt è uguale a un joule per secondo.) Nell'eruzione del Mount St. Helens gli $1,7 \times 10^{18}$ joule sono stati liberati nell'arco di nove ore, per una potenza media di circa 5×10^{13} watt ($1,7 \times 10^{18}$ diviso per 32 400 secondi). La potenza di emissione del vulcano si potrebbe dunque paragonare alla sequenza esplosiva di 27 000 bombe del tipo di Hiroshima: quasi una ogni secondo per nove ore.

Negli ultimi 4500 anni il Mount St. Helens ha avuto almeno 20 periodi eruttivi, meritandosi dagli indiani dell'America nordoccidentale l'appellativo di «Loowit», ossia Signora del fuoco. Prima delle eruzioni del 1980 il Mount St. Helens era stato attivo l'ultima volta fra il 1831 e il 1857. È uno dei 15 maggiori vulcani della Catena delle Cascate, che dal Lassen Peak in California si estende verso nord fino al Mount Garibaldi nella Columbia Britannica. Le Cascate fanno parte della «cintura di fuoco», ovvero delle catene vulcaniche che cingono da vicino il Pacifico. Le montagne sono spinte verso l'alto in corrispondenza delle zone di subduzione là dove le zolle tettoniche in movimento del bacino del Pacifico s'immergono al di sotto delle zolle contigue a ovest, a nord e a est. La subduzione delle zolle del Pacifico provoca grandi terremoti e fornisce il magma e la pressione che alimentano la potenza dei vulcani della cintura di fuoco. In periodo storico sono stati attivi lungo il margine del Pacifico circa 400 vulcani.

La frequenza dell'attività vulcanica sembra in parte dipendere dalla velocità con cui le zolle in subduzione s'immergono al di sotto delle zolle contigue. In Indonesia e in Giappone, dove la velocità annua di subduzione è pari a sei o sette centimetri, si ha - solitamente - almeno un'eruzione vulcanica all'anno. La minore frequenza di eruzioni nelle Cascate sembra doversi attribuire alla minore velocità di convergenza della

I modelli grafici (elaborati con il calcolatore) del Mount St. Helens, visto da nord-est prima *(in alto)* e dopo *(in basso)* la violenta eruzione del 18 maggio 1980 dimostrano come l'altezza della montagna sia diminuita di 400 metri, lasciando un cratere profondo 750 metri. I modelli sono stati elaborati nell'ambito del programma Digital Elevation Model presso il Western Mapping Center dell'US Geological Survey.

Carta geologica della regione devastata dall'eruzione del 18 maggio. L'area in colore scuro indica i depositi piroclastici, costituiti da minuti frammenti di rocce. L'area a tratteggio è quella caratterizzata dalla caduta di cenere vulcanica. L'area in colore medio chiaro è quella in cui gli alberi furono abbattuti da un vento impetuoso. Nella zona in colore chiaro gli alberi rimasero in piedi, ma i loro aghi furono distrutti dal calore. I depositi di valanga sono contrassegnati dal colore grigio chiaro, le colate di fango dal grigio scuro.

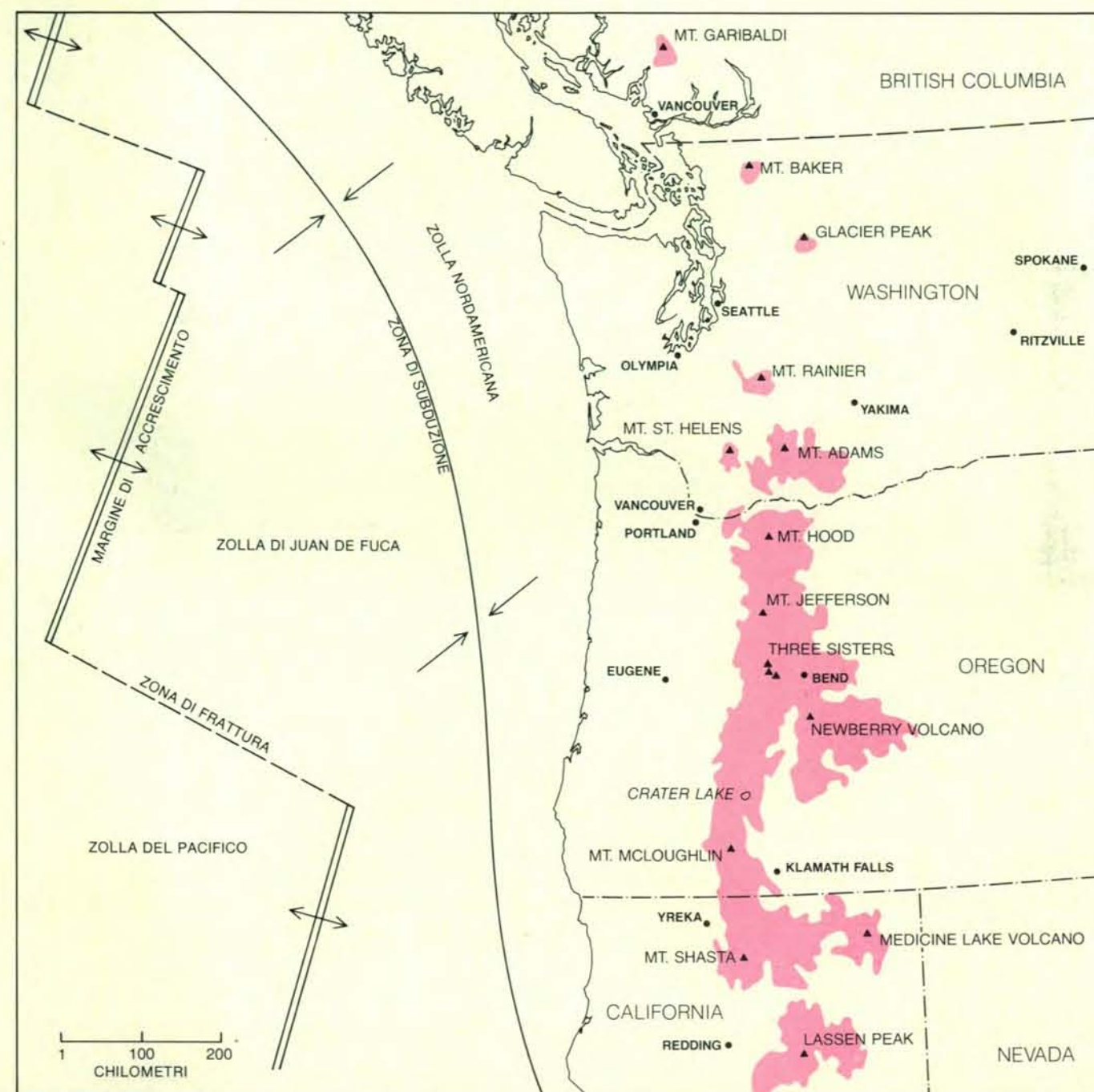
zolla Nord-americana e della zolla di Juan de Fuca, che è immediatamente a ovest: due o tre centimetri all'anno.

All'inizio degli anni sessanta Dwight Crandell e Donal Mullineaux dell'US Geological Survey intrapresero uno studio per valutare i rischi relativi ai vulcani quiescenti, ma potenzialmente attivi della Catena delle Cascate. Essi inaugurarono il loro programma con un'indagine sui depositi vulcanici prossimi al Mount Rainier, 75 chilometri a nord-nord-est del Mount St. Helens. Il Mount Rainier aveva dato manifestazioni eruttive l'ultima volta tra il 1820 e il 1854. Basandosi sulle datazioni - ottenute con il metodo del carbonio 14 - di eruzioni più antiche e seguendo l'asserto dei geologi secondo cui ciò che accadde nel passato può nuovamente verificarsi, Crandell e Mullineaux poterono formulare una previsione di massima sul rischio potenziale di ognuno dei vulcani presi in esame.

Il rapporto sul Mount St. Helens fu pubblicato dai due studiosi nel 1978. Essi concludevano che il Mont St. Helens aveva un brutto passato: infatti negli ultimi 4500 anni era stato il più attivo e il più esplosivo fra tutti i vulcani presenti nel territorio relativo ai 48 stati contigui. In tale periodo il vulcano produsse duomi di lava tanto viscosa da non poter defluire, grandi depositi di cenere vulcanica contenenti parecchi blocchi di pomice, colate di materiale piroclastico (frammenti incandescenti che l'attività eruttiva sminuzza e fluidizza), colate di lava e ingenti colate di fango nelle valli fluviali irradiantisi dal vulcano. L'intervallo medio tra i periodi eruttivi era stato di 225 anni.

Sulla base del loro studio Crandell e Mullineaux dichiaravano: «In futuro, probabilmente, il Mount St. Helens erutterà con la stessa violenza e intermittenza manifestate nel recente passato geologico, e le future eruzioni influiranno sull'attività e la sicurezza dell'uomo, sulle proprietà, sull'agricoltura e in generale sul benessere economico di una vasta area. Il comportamento del vulcano suggerisce che l'attuale intervallo di quiescenza sia inferiore a un millennio: è anzi probabile che si verifichi un'eruzione entro i prossimi 100 anni, forse proprio prima della fine di questo secolo.»

Molte eruzioni vulcaniche sono precedute da «sciami» di piccoli terremoti. Benché non tutti i vulcani della Catena delle Cascate siano controllati con sismografi, fortunatamente lo era il Mount St. Helens. Infatti l'Università di Washington teneva un sismografo sul fianco occidentale del vulcano, collegato a distanza con Seattle. Il 20 marzo 1980, alle ore 3 e 47 pomeridiane secondo il Pacific Standard Time si verificò sotto la montagna un terremoto di magnitudo 4 della scala Richter. Quando l'insolito evento fu seguito da un numero crescente di terremoti locali, apparve chiaro che si stava progressivamente sviluppando un più intenso sciame di terremoti. Furono installati altri sismografi per migliorare le registrazioni e rendere più precisa la localizzazione delle scosse.

Il 25 marzo l'energia sismica liberata dallo sciame raggiunse il suo valore più alto: 47 terremoti di magnitudo uguale o superiore a 3 si verificarono in un periodo di 12 ore a piccola profondità sotto il fianco settentrionale della montagna. Le prime piccole esplosioni di vapore incominciarono due giorni più tardi alle 12 e 36, dando luogo a un nuovo cratere ampio circa 70 metri sulla sommità innevata e ghiacciata. Si formarono inoltre - con orientazione est-ovest - grandi fenditure nella neve e nel ghiaccio, che rivelavano che nella zona sommitale veniva costituendosi un blocco inferiormente fagliato. Un secondo cratere si aprì il 29 marzo. Fiamme bluastre, probabilmente brucianti acido solfidrico, furono osservate dall'alto durante la notte. Il

Carta tettonica che mostra la relazione fra le zolle nordamericana, del Pacifico e di Juan de Fuca a ovest. In corrispondenza del margine di subduzione, la zolla di Juan de Fuca s'immerge sotto quella nordamericana dando origine ai vulcani della Catena delle Cascate *(triangolini neri)*. Le aree in colore rappresentano depositi vulcanici relativi agli ultimi due milioni di anni. I dati sono dell'US Geological Survey.

30 marzo si rilevarono novantatre piccole eruzioni di vapore e di cenere.

Il primo aprile i sismografi registrarono l'inizio del tremore vulcanico, ossia quella vibrazione più o meno continua del suolo che già era stata osservata in molti vulcani attivi. Di tale tremore non si conosce precisamente la causa, ma è presumibile che esso rifletta il movimento del magma o il gorgogliare dei gas in esso contenuti.

La nuova fase eruttiva - la prima nella Catena delle Cascate dopo la cessazione nel 1917 dell'attività del Lassen Peak - era divenuta un'attrattiva per i curiosi. Nolan Lewis, direttore dei Servizi di emergenza per la Contea di Cowlitz, Washington, in cui si trova il Mount St. Helens, riferì che «domenica [30 marzo], mentre l'atmosfera era tersa, le strade sulla montagna assomigliavano a quelle di Seattle nelle ore di punta». Quando si fece via via più probabile il pericolo di eruzioni più violente, il governatore Dixy Lee Ray ordinò che fossero aumentati i posti di blocco sulle strade di accesso.

Le piccole eruzioni di vapore e di cenere continuarono, alcune sotto forma di esplosioni singole, altre come getti intermittenti della durata di qualche ora. Colonne di cenere e di vapore s'innalzavano fino a 3 chilometri sopra la cima. Il cratere si sviluppò fino ad assumere la forma di una depressione ovale di ampiezza pari a 500 metri per 300 con una profondità di 200 metri. La cenere consisteva di frammenti di vecchia roccia vulcanica; le emissioni di gas comprendevano piccole quantità di anidride carbonica, anidride solforosa, acido solfidrico e acido cloridrico, insieme a grandi quantità di vapore. Le piccole eruzioni esplosive derivavano evidentemente dall'acqua freatica presente nella parte alta del cono vulcanico, riscaldata al di sopra del punto di ebollizione sotterraneo e prontamente trasformata in vapore, così come accade in un geyser, ma con una energia sufficiente a incorporare particelle di cenere e a proiettarle fuori dal cratere. L'energia liberata dalle esplosioni di vapore fino al 18 maggio è stata stimata in 10^{14} joule.

Oltre al protrarsi dell'elevato livello di attività sismica (circa 50 terremoti al giorno, di magnitudo uguale o superiore a 3) vi era un altro segno preoccupante. Fin dal 27 marzo numerosi osservatori avevano notato che lo spostamento del blocco alla sommità della montagna sembrava essere accompagnato da una protuberanza, ovvero da un sollevamento, della parte superiore del fianco settentrionale del cono vulcanico. La protuberanza continuò a crescere nei primi giorni di aprile manifestandosi attraverso un sistema di grosse fenditure che si estendevano via via nella coltre di ghiaccio e di neve. Carte fotogrammetriche costruite dall'US Geological Survey in base a fotografie aeree riprese il 12 aprile rivelarono che la protuberanza aveva un diametro di circa 2 chilometri e si era innalzata per almeno 100 metri.

I rilievi a terra effettuati tra la fine di aprile e i primi di maggio 1980 stabilirono che la protuberanza si espandeva in modo continuo verso nord - più o meno orizzontalmente - alla velocità di circa 1,5 metri al giorno. L'area soggetta a rapida deformazione si trovava al centro della zona sismica, 2 chilometri sotto. Gli osservatori presenti erano concordi nel ritenere che il continuo sciame sismico e la maggiore deformazione superficiale fossero una prova evidente che il magma stava risalendo verso la superficie. Si aspettavano quindi un'eruzione più violenta oppure una valanga sul lato settentrionale in espansione. Gli unici interrogativi riguardavano il momento in cui ciò si sarebbe verificato e l'entità del fenomeno.

Robert Christiansen, il ricercatore addetto al programma di controllo dell'US Geological Survey, riesaminò la sequenza storica dell'attività di altri vulcani, in modo particolare del Lassen Peak, unico vulcano della Catena delle Cascate con eruzione bene osservata, e del Bezymianny sulla costa pacifica dell'URSS, nella penisola di Kamchatka, esploso con violenza nel 1956. Christiansen concluse che per il Mount St. Helens era più verosimile pensare a una eruzione del tipo di quella del Lassen nel 1915. Non si poteva comunque escludere un'eruzione della medesima intensità di quella del Bezymianny. Una terza supposizione era che l'attività potesse decrescere senza dar luogo a un'eruzione più consistente.

Intanto, il 7 maggio, dopo circa due settimane di modesta attività visibile, ripresero piccole esplosioni di vapore e di cenere. Benché l'attività sismica fosse proseguita senza variazioni e la protuberanza fosse costantemente aumentata, l'assenza di una drammatica attività visibile indusse gli abitanti della regione a chiedere all'autorità politica ragione del blocco della zona circostante la montagna. Il 15, 16 e 17 maggio l'attività sismica e il rigonfiamento continuarono senza, tuttavia, indizi di vapore e di cenere.

Il 18 maggio, alle 7 antimeridiane secondo il Pacific Daylight Time, Dorothy e Keith Stoffel, due geologi dello stato di Washington, s'imbarcarono su un aereo da turismo all'aeroporto di Yakima prossimo al Mount St. Helens, intraprendendo il loro primo volo di ricognizione. Effettuarono diversi passaggi attorno e sopra il vulcano. Nonostante la mattinata fosse limpida e luminosa, non si scorgeva alcuna attività. Alle 8 e 32 la montagna fu scossa da un terremoto di magnitudo 5,1 il cui fuoco si trovava sotto il fianco settentrionale del monte. In quel momento gli Stoffel si trovavano proprio sulla verticale della cima, che osservavano dall'altezza di 400 metri. Notarono molte piccole frane di ghiaccio precipitare lungo i fianchi più ripidi del cratere. Quindici secondi più tardi i due geologi erano i più prossimi testimoni dell'inizio di un'immensa eruzione vulcanica innescata da una delle più grandi frane della storia.

«L'intero fianco settentrionale della sommità del cratere cominciò a muoversi repentinamente come un'enorme massa unica» ricordava Dorothy Stoffel. «L'intera massa cominciò a incresparsi e a ribollire senza spostarsi lateralmente. Poi tutto il versante nord della cima si mise in movimento verso settentrione lungo un piano di scivolamento profondamente incavato.»

Qualche secondo dopo avvenne una grande esplosione. È singolare che gli Stoffel non l'udissero affatto benché si trovassero proprio a est della cima. Dalla loro posizione la nube dell'esplosione iniziale assomigliava a un ombrello inclinato verso nord: indi precipitò lungo il pendio. La loro prima preoccupazione fu quella di salvarsi. Benché si tuffassero a tutta velocità, la nube grigia in espansione stava per raggiungerli: la sfuggirono infine virando a sud. Dietro di loro, gigantesche nubi di cenere ribollivano verso l'alto, spingendosi verso nord e verso nord-ovest. A oriente le nubi si espandevano a forma di ombrelli fluttuanti, illuminate da abbaglianti saette lunghe migliaia di metri. Mezz'ora più tardi gli Stoffel atterravano a Portland.

Dall'aria essi avevano osservato più di due chilometri cubi di roccia e di ghiaccio frantumati che precipitavano nello Spirit Lake e nel ramo settentrionale del Toutle River. Fluidificata dal vapore esplodente, la valanga accelerava rapidamente fino a una velocità di 250 chilometri all'ora. Un lobo della gigantesca massa attraversava il braccio occidentale dello Spirit Lake, proseguendo verso nord nella valle ulteriore. Un lobo vicino si muoveva rapidamente attraverso un'altra valle con una tale quantità di moto per cui poté risalire un colle alto 360 metri che limitava la valle a nord. La maggior parte dei detriti fluidizzati s'incanalava lungo la valle di Toutle, formando un deposito ondulato lungo 21 chilometri, ampio da 1 a 2 chilometri e profondo fino a 150 metri. L'energia gravitazionale della valanga era di circa 5×10^{16} joule (circa l'equivalente di 12 megaton).

Poiché la grande valanga di roccia e di ghiaccio aveva provocato una subitanea diminuzione della pressione entro il cono vulcanico, l'acqua freatica surriscaldata si trasformava immediatamente in vapore. Nel contempo i gas disciolti esplodevano dal corpo magmatico superficiale, recentemente intruso nella zona superiore della montagna. L'esplosione del vapore, quella del magma e l'immane valanga si unirono a determinare una violenta eruzione laterale di nubi di vapore ardenti (fino a 300 gradi centigradi), dense, pregne di detriti, le quali si sollevavano verso nord - dal fianco squarciato del monte - a velocità variabile da 100 a 400 chilometri all'ora. L'eruzione di vapore e il suo carico fluidizzato di frammenti di roccia vulcanica devastarono 550 chilometri quadrati di terreno montano a nord-ovest, a nord e a nord-est del Mount St. Helens. Le dense nubi, aderendo al suolo, rotolarono sopra quattro dei maggiori rilievi e delle valli interposte, spingendosi fino a 28 chilometri di distanza.

La distruzione fu completa. Nel raggio dei primi chilometri, interi alberi del diametro da uno a due metri erano stati sradicati e spazzati via dalla nube, resa torbida dall'esplosione. Più in là, vi era una zona, ampia da 10 a 15 chilometri, nella quale un vento rabbioso aveva spezzato come fossero fiammiferi stupendi esemplari di abete di Douglas. Ai margini dell'area distrutta gli alberi erano ancora in piedi, ma i loro aghi erano irrimediabilmente bruciacchiati.

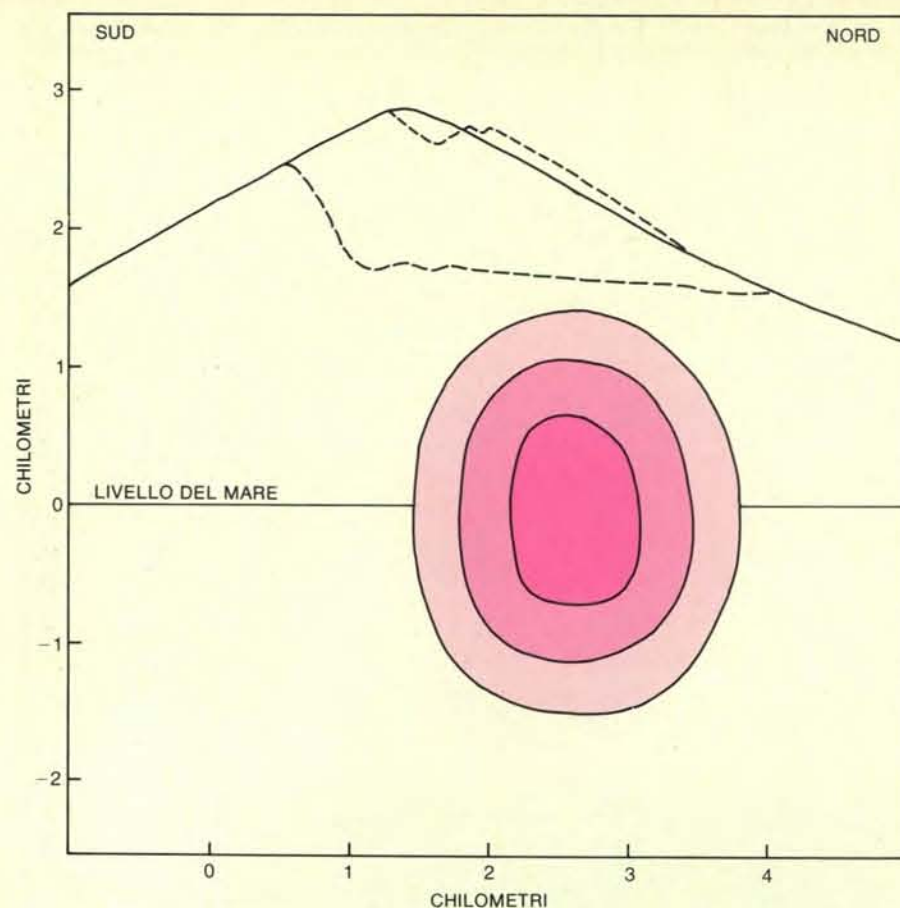
Le fotografie del Mount St. Helens, prima e dopo l'eruzione, documentano l'ampiezza del nuovo cratere. La fotografia in alto era stata eseguita da uno degli autori (Robert Decker) nel giugno 1970. L'altitudine della cima (vista qui da nord-nord-est) era di 2950 metri; si elevava da un basamento la cui altitudine è di 1000 metri. La fotografia in basso è stata scattata, invece, da Ray Foster dei Sandia Laboratories nel luglio del 1980. Il cratere (visto qui da nord) presenta un diametro di 2 chilometri. L'altitudine dell'orlo del cratere è di circa 2400 e 2550 metri, il fondo del cratere si trova a un'altitudine compresa tra i 1800 e i 1900 metri. Colate piroclastiche coprono buona parte del primo piano.

La prima impressione che si aveva osservando la zona colpita dalla veemenza del vento era che una grande onda d'urto o fronte di concussione avesse investito gli alberi propagandosi radialmente verso il basso dalla vetta esplodente. Ciò, comunque, non fu dimostrato. I superstiti, vicino ai bordi dell'area devastata, avevano udito soltanto un'esplosione non molto forte ovvero una specie di fragore, due o tre minuti prima che la nube nera scendesse su di loro accompagnata da caldi venti d'uragano. La velocità del fronte di nubi prodotte dall'eruzione di vapore era molto inferiore alla velocità del suono. A un'ispezione più attenta l'aspetto degli alberi abbattuti denotava l'azione di moti vorticosi e turbolenti. Ai confini dell'area devastata gli alberi mostravano di essere stati investiti da un vento diretto verso le vallate sottostanti, anche dove apparivano piegati verso il punto di origine delle nubi che si sollevavano.

Evidentemente la gravità aveva fornito energia alla densa massa fluidizzata, mentre andava decrescendo l'energia esplosiva dell'eruzione iniziale di vapore. Successivamente, attenuandosi la turbolenza dei venti, le ceneri e i frammenti di roccia che erano in sospensione dentro le nubi formarono depositi superficiali il cui spessore decresceva - da circa un metro a un centimetro - con l'aumentare della distanza dalla sorgente. Frammenti spigolosi di diametro superiore a parecchie decine di centimetri, relativi a rocce vulcaniche sia vecchie sia recenti (queste più calde), furono trasportati dalle nubi fino a distanze di 10-15 chilometri. Sul versante dell'eruzione gli alberi si rivelavano carbonizzati e anneriti fino a una distanza di oltre 7 chilometri a nord e a nord-ovest, e di 18 chilometri a nord-est.

Il carattere fluidizzato delle masse clastiche proiettate dall'esplosione risulta ancor più evidente nei depositi formatisi sui versanti scoscesi. Qui, dopo il deposito iniziale, flussi secondari hanno dato luogo alla formazione di depositi stagnanti, spessi decine di metri, nei bacini e sul fondo delle valli. Il volume totale dei sedimenti prodotti dall'esplosione è di circa 18 000 milioni di metri cubi. Di questi circa 60 milioni sono di origine magmatica recente. L'energia termica fornita dalla componente magmatica è stata approssimativamente di 2×10^{17} joule.

L'acqua freatica surriscaldata nell'interno del vulcano ha costituito probabilmente un'altra sorgente di energia per la violenta esplosione di vapore. Supponendo una porosità del 15 per cento in una massa rocciosa del volume da 2000 a 3000 milioni di metri cubi, impregnata d'acqua alla temperatura di 175 gradi centigradi, l'acqua medesima produrrebbe un'energia aggiuntiva di 10^{17} joule. Probabilmente il vapore della nube eruttiva si è generato in due modi. Il calore dell'acqua freatica erompente avrebbe generato $4,4 \times 10^{10}$ chilogrammi di vapore, mentre il calore addizionale contenuto nei 60 milioni di metri cubi di frammenti magmatici avrebbe trasformato in vapore altri $8,8 \times 10^{10}$ chilogrammi d'acqua. Alla temperatura di 100 gradi centigradi a pressione atmosferica, la quantità complessiva di vapore ($1,3 \times 10^{11}$ chilogrammi) si sarebbe espansa fino a raggiungere il volume di 220 miliardi di metri cubi.

Immagini verticali del Mount St. Helens, prima e dopo l'eruzione, ottenute rispettivamente il primo maggio *(a sinistra)* e il 19 giugno *(a destra)* da un aereo U-2 della National Aeronautics and Space Administration. La cima è in basso a destra. Le immagini, riprese con negativo all'infrarosso, sono state stampate in falso colore. Le aree rosse corrispondono al verde della vegetazione (per lo più abeti di Douglas). Nella fotografia a destra, le aree in grigio sono quelle devastate dall'eruzione del 18 maggio; sulla cima si osserva una nube di fumo elevantesi dal cratere.

I profili topografici rappresentano il Mount St. Helens nell'agosto 1979 *(linea continua in alto)*, il primo maggio 1980 quando sul fianco nord si formò una protuberanza *(linea tratteggiata superiore)* e primo luglio 1980 *(linea tratteggiata inferiore)*. L'area in colore sotto i profili rappresenta la zona di origine delle migliaia di sismi verificatisi fra il 20 marzo e il 18 maggio 1980. Il colore più scuro indica l'area di maggiore addensamento degli ipocentri. I dati sono stati forniti dall'US Geological Survey e dal Department of Geophysics dell'Università di Washington.

La massa di acqua freatica calda, che si è supposta presente nel vulcano ($3,75 \times 10^{11}$ chilogrammi), è quasi tre volte superiore a quella che sarebbe necessaria per ottenere il volume di vapore calcolato. I valori relativi all'energia e al volume del vapore sono soltanto approssimativi, benché appaiano ragionevoli in rapporto all'area devastata - 550 chilometri quadrati - dalle nubi prodotte dall'eruzione laterale. L'ultimo deposito eruttivo era costituito da uno strato di cenere umida, profondo più di sei centimetri, contenente in molti punti sfere di fango del diametro di un pisello, che i vulcanologi chiamano «lapilli di accrescimento». I lapilli si erano formati intorno a nuclei di gocce di pioggia condensatisi dalle nubi di vapore.

Alle 9 antimeridiane del 18 maggio la fase parossistica dell'eruzione era superata, nonostante il persistente rumoreggiare, per buona parte della giornata, di una colonna verticale (di gas e di cenere), che raggiungeva altezze superiori ai 20 chilometri: infine - intorno alle 5 e 30 pomeridiane - i fenomeni cominciarono a decrescere. La colonna di gas e di cenere, in esplosione pressoché continua e in movimento verso l'alto, era determinata dal degasarsi della parte superficiale del corpo magmatico in graduale risalita dalle profondità. L'azione abrasiva del flusso ascendente continuò ad allargare la forma a ferro di cavallo assunta dal cratere, la quale era stata inizialmente causata dalla valanga e dalle esplosioni laterali.

I venti d'alta quota soffiarono tutto il giorno verso nord-est e la cenere cominciò a cadere - a metà della mattinata - sulle città della zona centrale dello stato di Washington. A Yakima, lontana 150 chilometri, la prima cenere caduta formò uno strato di «sale e pepe», consistente in frammenti di roccia scura e in cristalli di feldspato lievemente colorato, ognuno della dimensione di un granello di sabbia. Sopra questo strato se ne depositò uno più sottile, di vetro vulcanico, le cui particelle avevano le dimensioni proprie del silt. Trenta chilometri a nord di Yakima la coltre di cenere raggiungeva lo spessore di circa 20 millimetri. A est lo strato della cenere minuta era più spesso, superando i 70 millimetri presso Ritzville (Washington) a 330 chilometri dal Mount St. Helens: la cenere sembrava polvere di talco.

Presso Spokane (Washington), 430 chilometri a nord-est del vulcano, lo spessore della cenere era di soli 5 millimetri, ma dalle 3 pomeridiane la visibilità andò riducendosi fino a tre metri sul calar della sera. Verso il mezzogiorno del 19 maggio cadde della cenere su Denver. In meno di tre giorni la nube aveva attraversato gli Stati Uniti. Il peso della cenere caduta era equivalente a 150 milioni di metri cubi di magma e rappresentava un dispendio di energia termica dell'ordine di 5×10^{17} joule. Gli studi inerenti alle tracce di cenere caduta al di fuori dell'area suddetta e agli aerosol ancora sospesi nella stratosfera, costituiti da cenere molto fine e da acido solforico, indicano che si era disperso un ulteriore volume di magma, pari a 100 milioni di metri cubi.

Sulla base dell'esperienza di altre eruzioni, che avevano immesso polvere e aerosol nella stratosfera, si può dedurre che le particelle molto fini impiegheranno uno o due anni per depositarsi. L'effetto sul clima mondiale, se mai ci sarà, non è ancora evidente.

Inondazioni e colate di fango furono un altro dei principali aspetti dell'eruzione. Le colate consistevano di una fanghiglia di cenere vulcanica e di minute particelle di roccia mescolate ad acqua, aventi una consistenza simile a quella del cemento bagnato. Il materiale solido derivava dalla coltre di cenere depositatasi intorno alla montagna e dalla roccia frantumata del deposito di valanga; è probabile che l'acqua eccedente avesse diverse origini: neve e ghiaccio disciolti, acqua dello Spirit Lake e del ramo settentrionale del Toutle River spostata dalla valanga, acqua derivante dal sistema idrotermale sconvolto, non trasformatasi istantaneamente in vapore o in vapore condensato.

Il fronte della prima colata di fango arrivò nel ramo meridionale del Toutle River presso Silver Lake alle 10 e 50 antimeridiane del 18 maggio, superando di 30 centimetri il più alto livello di piena registrato in epoca storica. La colata più consistente venne dal ramo settentrionale del Toutle River e raggiunse il livello più alto intorno alle 19, provocando la distruzione della stazione di misura presso il Silver Lake. Le tracce lasciate dal fango e dall'acqua indicavano che i livelli storici di piena erano stati superati di nove metri.

A valle del Toutle i depositi ostruirono i canali del Cowlitz River e provocarono gravi intasamenti del canale di navigazione del Columbia River. Il volume dei depositi di fango era di circa 100 milioni di metri cubi: un volume pressappoco uguale di acqua sarebbe stato coinvolto nella fluidificazione delle colate di fango.

Qualche tempo dopo la caduta della valanga iniziale e l'esplosiva eruzione di vapore, cominciarono a precipitarsi lungo il versante nord del Mount St. Helens, attraverso la breccia del nuovo cratere, colate piroclastiche di cenere minuta e di blocchi di pomice. Queste emulsioni fluidizzate, costituite da frammenti di roccia e di vetro incandescenti frammisti a caldi gas vulcanici, fuoriuscivano dal cratere (perché più dense) alla base della nube di cenere che andava sollevandosi. Le successive colate si riversarono sul pendio settentrionale a velocità superiori ai 1000 chilometri all'ora, coprendo la frana iniziale e i depositi dell'esplosione, fino a raggiungere il limite meridionale dello Spirit Lake. Quando le colate piroclastiche calde (300-370 gradi centi-

Il momento dell'eruzione del 18 maggio (esattamente 20 secondi dopo l'inizio del fenomeno eruttivo) fu colto da Keith e Dorothy Stoffel, due geologi che stavano volando in quel momento sulla montagna. L'eruzione fu preceduta dalla valanga. Gli Stoffel sono riusciti a fuggire.

Un duomo di lava si formò nel cratere fra il 13 e il 20 giugno. Aveva un diametro di 300 metri e un'altezza di 65; fu poi distrutto dall'eruzione successiva del 22 luglio. Questa fotografia aerea è stata scattata da Maurice e Katia Krafft del Centre de Volcanologie di Cernay (Francia).

gradi) ebbero raggiunta l'acqua, provocarono esplosioni secondarie, facendo innalzare nubi di cenere e vapore fino all'altezza di 2000 metri. Le colate di cenere e di pomice continuarono fino alla sera del 18 maggio: il volume totale era di 250 milioni di metri cubi e l'energia termica stimata pari a $3,3 \times 10^{17}$ joule.

Il costo in vite umane fu relativamente grande: risultarono morte o disperse 62 persone. I danni all'economia, soprattutto per l'industria del legname, superarono il miliardo di dollari. Ma il danno più grave, forse, fu di carattere psicologico. Negli Stati Uniti nordoccidentali i vulcani della Catena delle Cascate, che venivano considerati come guardiani silenziosi, costituiscono ormai una minaccia latente.

L'enorme valanga e l'eruzione del 18 maggio furono seguite da minori esplosioni di cenere il 25 maggio, il 12 giugno, il 22 luglio, il 7 agosto e il 16-18 ottobre. Il 13-20 giugno, l'8-9 agosto e il 18-19 ottobre, furono estrusi all'interno del cratere duomi di lava viscosa. Il volume totale di magma effuso nel periodo intercorrente fra il 25 maggio e il 19 ottobre fu di circa 50 milioni di metri cubi, con un contenuto di energia valutato intorno a $1,7 \times 10^{17}$ joule (un ordine di grandezza inferiore a quello dell'eruzione del 18 maggio). L'eruzione del 25 maggio si verificò di notte e in condizioni di tempo non buone: il suo inizio coincise probabilmente con l'improvviso incremento del tremore vulcanico alle 2 e 38 antimeridiane. Alle 2 e 45 il radar del National Weather Service indicava che la colonna di cenere aveva raggiunto un'altezza di 14 000 metri: diminuì in altezza entro un'ora, benché il vulcano avesse continuato per tutto il giorno a emettere nubi di cenere più piccole. Si ebbero anche nuove colate piroclastiche di cenere e di blocchi di pomice le quali ricoprirono parte della medesima area - sul fianco settentrionale del Mount St. Helens - che già era stata percorsa dalle colate piroclastiche iniziali. Benché l'eruzione di cenere fosse molto meno voluminosa di quella del 18 maggio, la direzione del vento era più variabile così che un sottile strato di cenere cadde su vaste zone dello stato di Washington occidentale e dell'Oregon, compresa l'area metropolitana di Portland.

Il 12 giugno le esplosioni di cenere furono simili all'eruzione del 25 maggio. Il tremore vulcanico cominciò nel pomeriggio: la prima emissione di cenere si verificò alle 17 e 05, raggiungendo un'altezza di 4000 metri. Un'eruzione di cenere molto più grande incominciò alle 21 e 09, innalzandosi fino a 15 000 metri. Il fenomeno si esaurì in breve tempo dopo la mezzanotte. Il giorno successivo, osservazioni dall'elicottero rivelarono che altre colate piroclastiche, di uno spessore variabile da 2 a 10 metri e con temperature superiori a 600 gradi centigradi, erano discese verso lo Spirit Lake. Un duomo di lava cominciò a formarsi nel cratere di esplosione dopo l'eruzione del 12 giugno: quando fu rilevato il 15 giugno, il duomo era largo 200 metri e alto 40, solcato da grandi crepe incandescenti. Continuò a sollevarsi di circa 6 metri al giorno: il 20 giugno aveva già raggiunto l'altezza di 65 metri.

Il vulcano rimase tranquillo fino al 22 luglio, quando, il mattino, incominciarono piccoli terremoti superficiali sotto il cratere. Le scosse aumentarono di numero durante la giornata, ma non si registrò alcun tremore vulcanico. Improvvisamente alle 17 e 14 di una limpida giornata estiva cominciò a levarsi dalla montagna una grande nube di cenere: il radar ne registrò la sommità all'altezza di 14 000 metri. Una seconda nube di cenere eruttò alle 18 e 25, raggiungendo un'altezza di 18 000 metri in appena sette minuti e 23 secondi, e con una velocità media di 2,2 chilometri al minuto. Il terzo e più persistente getto di cenere ebbe inizio alle 19 e 01 e durò più di due ore, elevandosi fino a 14 000 metri.

Benché i geologi e gli aerei del Forest Service si fossero allontanati dopo l'allarme provocato dall'intensificarsi dei lievi terremoti superficiali, le eruzioni furono osservate da elicotteri e da altri velivoli. Colate piroclastiche fluirono dalla bocca eruttiva durante la seconda e la terza esplosione di cenere, spargendosi lungo il pendio settentrionale del vulcano in direzione dello Spirit Lake. Richard Hoblitt dell'US Geological Survey narrò come testimone oculare:

«Stavamo volando da est a ovest circa un chilometro e mezzo a nord della bocca eruttiva quando iniziò la seconda eruzione. Dopo un periodo della durata di pochi secondi, durante il quale era aumentata la velocità di emissione dei gas, si formò una fontana di cenere alta circa 500 metri. Le proiezioni della fontana, inarcandosi al sommo e ricadendo in vicinanza della bocca eruttiva, diedero origine a una colata piroclastica che si mosse rapidamente verso nord all'esterno dell'anfiteatro craterico. Ci allontanammo verso occidente il più presto possibile». Poche persone hanno potuto osservare una colata piroclastica così da vicino e sono sopravvissute per descriverla.

Lo spessore delle nuove colate di cenere e di blocchi di pomice variava da uno a due metri. Se ne misurò la temperatura il giorno seguente: il massimo, 705 gradi centigradi, si rivelò alla profondità di un metro e mezzo. Il 22 luglio le nubi di cenere andarono alla deriva verso nord-est: soltanto lievi cadute di cenere furono segnalate nella zona centrale e in quella orientale dello stato di Washington.

L'eruzione del 22 luglio era stata preceduta da un cambiamento del modo di emissione dei gas. Un altro cambiamento si verificò ai primi di agosto. Il tremore vulcanico riprese il 7 agosto, dopo mezzogiorno. Ammoniti da questi due segni, gli investigatori evacuarono la zona pericolosa. Alle 16 e 26 ebbe inizio un'eruzione, che rapidamente generò una colonna di cenere alta 13 000 metri. Piccole colate piroclastiche si misero in rapido movimento sul pendio sottostante allo squarcio del fianco settentrionale della montagna. Minori eruzioni di cenere persistettero nel tardo pomeriggio e nella serata: un'altra violenta esplosione si verificò alle 22 e 32. L'8 e il 9 agosto si formò un nuovo duomo di lava entro il cratere.

Per più di due mesi prevalse la quiete: il 16 settembre avvennero lievi terremoti, simili a quelli che avevano preceduto l'eruzione del 22 luglio: aumentando il loro numero, nella prima serata fu proclamato lo stato di allarme. Seguì un'eruzione con inizio alle 21 e 58. Nei due giorni successivi, quattro esplosioni di cenere proiettarono nubi fino a un'altezza di 14 000 metri, distruggendo il duomo di lava formatosi nell'agosto. Alcune eruzioni furono accompagnate da modeste colate di materiale piroclastico che discesero lungo il fianco settentrionale del monte. Ceneri di colore chiaro caddero a sud e a sud-est, rammentando all'area metropolitana di Portland che il Mount St. Helens era ancora attivo. Il 18-19 ottobre si sviluppò nuovamente un duomo di lava, ancora più imponente.

Il grado di esplosività delle eruzioni vulcaniche è molto vario. Alle Hawaii il getto aereo delle fontane di lava incandescente è spettacolare, ma innocuo e le colate di lava fluiscono lentamente lungo il pendio dalle bocche di emissione. Al contrario gigantesche esplosioni possono distruggere intere montagne quando una parte notevole dell'energia termica del magma si converte in energia meccanica. Un criterio per misurare il grado di esplosività dei vulcani è la natura dei loro prodotti. Le eruzioni effusive sono caratterizzate da colate laviche, le eruzioni esplosive da prodotti clastici: cenere vulcanica, frammenti e blocchi di roccia. Nelle Hawaii circa il 98 per cento dei prodotti eruttivi è rappresentato da lave. Nei vulcani che circondano il Pacifico la situazione è quasi antitetica: il 90 per cento circa dei materiali eruttivi è costituito da clasti.

I vulcani si trovano in tre diversi sistemi tettonici. In ognuno di tali sistemi i vulcani differiscono gli uni dagli altri secondo il grado di esplosività. Come si è visto, i vulcani della cintura di fuoco sorgono lungo le zone di subduzione là dove le zolle convergono. Tali vulcani sono generalmente esplosivi. Dove le zolle divergono i vulcani si accompagnano a *rift*: le eruzioni, specialmente quelle sottomarine, sono per lo più effusive. Tuttavia le eruzioni possono diventare esplosive quando si verifichino in acque poco profonde o attraverso la crosta continentale. I vulcani di «punto caldo» relativi alle zolle tettoniche sono in genere effusivi se si trovano sulla crosta oceanica (Hawaii) ed esplosivi quando sono situati sulla crosta continentale (Yellowstone).

I fattori preminenti che determinano il grado di esplosività dei vulcani sono la viscosità del magma, la concentrazione dei gas in esso disciolti, la quantità di acqua freatica dell'area craterica, la pressione

Gli sciami di scosse telluriche, la formazione della protuberanza, l'emissione di anidride solforosa e il tremore vulcanico, quattro tipi di fenomeni legati alle eruzioni, sono qui rappresentati in funzione del tempo. Le eruzioni principali sono indicate dalle linee verticali. La deformazione maggiore della protuberanza e l'energia sismica sono riferite al periodo antecedente all'eruzione del 18 maggio. I dati sono dell'US Geological Survey e del Department of Geophysics dell'Università di Washington.

superficiale. Le esplosioni vulcaniche sono provocate dalla rapida espansione dei gas che si trovano nel magma oppure a diretto contatto con esso. Non vi è liberazione di energia come in una bomba. Più persuasivo il confronto con l'esplosione di una caldaia a vapore. La viscosità del magma è paragonabile alla resistenza della caldaia: più alta la viscosità, più violenta è l'esplosione potenziale. La quantità di gas presente nel magma o a diretto contatto con esso è paragonabile al volume della caldaia. La differenza tra la pressione del gas nel magma e quella soprastante si può paragonare alla differenza di pressione fra l'interno e l'esterno della caldaia. Il vapore (acqueo) è uno dei principali gas magmatici, ma in un'esplosione vulcanica ha un ruolo importante anche l'anidride carbonica.

Consideriamo due esempi. Il primo concerne un magma basaltico poco viscoso con scarso contenuto in gas disciolti. Se questo magma risale alla superficie attraverso una frattura, l'anidride carbonica e l'acqua si separano gradualmente con il diminuire della pressione. Le bolle di gas si espandono, per la bassa viscosità, provocando l'espansione del magma spumoso. Tale processo è il più importante nel momento in cui il magma sta per raggiungere la superficie: dà origine, in corrispondenza della bocca eruttiva, a una fontana di lava incandescente. Se il medesimo magma basaltico scaturisse sotto una massa d'acqua marina profonda 2000 metri, la diminuzione della pressione non sarebbe sufficiente a separare il vapore acqueo dalla soluzione magmatica. Probabilmente si libererebbe in parte l'anidride carbonica, tuttavia l'espandersi del miscuglio di magma e di anidride carbonica sarebbe limitato e sgorgherebbe tranquillamente dalla bocca sottomarina. In acqua poco profonda, il grado di esplosività varia secondo la velocità con cui la lava incandescente si mescola all'acqua sovrastante.

Il secondo esempio ci è offerto da un magma più siliceo con una viscosità relativamente più elevata e con alto contenuto di gas disciolti. Se tale magma dacitico sale verso la superficie, i gas in esso disciolti cominciano a formare bolle. Tuttavia, per l'elevata viscosità del materiale fuso, le bolle in espansione non raggiungono uno stato di equilibrio: esse presentano una sovrapressione interna che è stimata superiore a parecchie centinaia di atmosfere. Quando si sia formato un certo numero di bolle e la pressione esterna sia sufficientemente ridotta, il magma si suddivide in brandelli che l'esplodere dei gas che si liberano espelle violentemente dalla bocca eruttiva. Il raffreddamento conseguente all'espansione dei gas determina la solidificazione dei brandelli magmatici, che vengono espulsi quali particelle ardenti, ma solide, di una emulsione. Tali miscugli fluidizzati (emulsioni) danno origine a nubi eruttive o a colate piroclastiche secondo la densità, la direzione e la velocità della loro eiezione.

Una delle eruzioni più simili a quella del Mount St. Helens fu quella del Bezymianny (in Kamchatka) nel 1955-56. Il vulcanologo sovietico G. W. Gorshkov ha diviso l'eruzione del Bezymianny in cinque fasi: 1) dal 29 settembre al 21 ottobre 1955, uno sciame di terremoti premonitori; 2) dal 22 ottobre al 30 novembre 1955, violente eruzioni di cenere; 3) dall'1 dicembre 1955 al 29 marzo 1956, modeste eruzioni di cenere con sollevamento di tutto il preesistente duomo di lava; 4) marzo 1956, gigantesca esplosione con eruzione laterale; 5) aprile 1956, formazione di un nuovo duomo di lava all'interno del cratere d'esplosione.

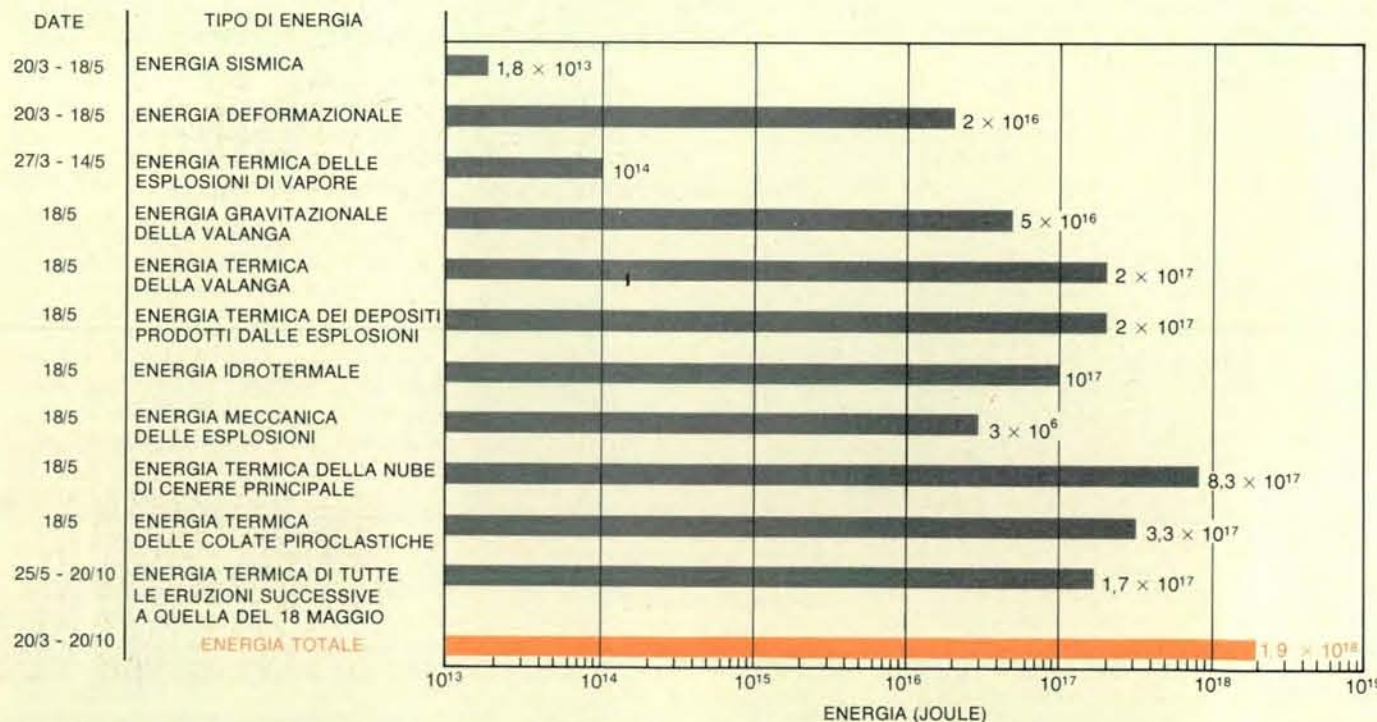
Per il volume di roccia spostata e per l'area coinvolta, l'eruzione del 30 marzo 1956 del Bezymianny è simile a quella del 18 maggio 1980 del Mount St. Helens. Queste furono le maggiori differenze: lo sciame di sismi, precedenti la prima eruzione di cenere, durò 23 giorni nel Bezymianny e sette nel Mount St. Helens; le eruzioni preliminari durarono 160 giorni nel primo vulcano e 53 nel secondo; nella prima fase eruttiva le maggiori esplosioni di cenere si verificarono nel Bezymianny. Nella più violenta eruzione del Bezymianny la maggior parte della roccia estrusa era costituita da nuovo magma, diversamente dalla valanga e dall'esplosione del 18 maggio nel Mount St. Helens, dove il magma di nuova formazione costituiva meno di un quinto della roccia estrusa. Forse non è privo di significato il fatto che il Bezymianny dal 1956 sia rimasto attivo. L'ultima intensa eruzione si è avuta nel 1979.

Il caso del Mount St. Helens ha rappresentato un valido test per le tecniche di previsione delle eruzioni vulcaniche. Non esiste una regola unica che consenta previsioni sicure: i fattori debbono essere valutati e poi interpretati alla luce dell'esperienza geologica. Fra tali fattori annoveriamo le statistiche relative alle eruzioni in tempi storici, la ricostruzione delle statistiche relative a eruzioni del periodo preistorico mediante i rilevamenti geologici e le datazioni. Le tecniche geofisiche coinvolte comprendono il controllo dell'attività sismica, delle deformazioni superficiali, dei campi elettrici e magnetici, delle temperature, in corrispondenza dei vulcani attivi o potenzialmente attivi. Le tecniche geochimiche riguardano il controllo del volume e della composizione dei gas, dei liquidi e dei solidi emessi dai vulcani. Con ripetute osservazioni dirette da terra o dall'aria si ricavano dati importanti sull'evoluzione dell'attività vulcanica. La maggior parte dei vulcani non è ancora stata sottoposta a queste indagini fondamentali, senza dire dei più sofisticati controlli strumentali.

Per il Mount St. Helens alcuni metodi si sono rivelati migliori di altri. I dati relativi all'attività in epoca storica documentano eruzioni nel periodo 1831-1857: tale campione, prescindendo dall'identificazione del Mount St. Helens come vulcano attivo, è troppo limitato per assumere un valore statistico. I rilevamenti e le datazioni geologiche hanno identificato 20 periodi eruttivi con prodotti diversi negli ultimi 4500 anni: si sono così potuti stabilire alcuni schemi importanti. Gli intervalli di quiescenza si distinguono in due «popolazioni»: da 100 a 150 anni e da 400 a 500 anni. Gli ultimi due intervalli di quiescenza prima del 1800 furono del tipo più breve. Anche i prodotti eruttivi dimostrano che le esplosioni di cenere e di clasti erano consuete e che le eruzioni preistoriche del Mount St. Helens interessavano vaste aree. Proprio questo tipo di analisi consentì a Crandell e a Mullineaux di prevedere la pericolosità del Mount St. Helens.

Nel 1980 il rilevamento della crescente attività sismica sotto il Mount St. Helens fu preludio alle prime piccole eruzioni esplosive verificatesi dopo una settimana. La fotogrammetria e le misurazioni elettro-ottiche di distanza documentarono la spettacolare deformazione superficiale conseguente all'accrescimento della protuberanza sul fianco settentrionale della montagna. Immediatamente prima dell'eruzione del 18 maggio non ci fu alcun cambiamento nel tasso di sismicità o in quello della deformazione superficiale, ma il protrarsi di tali fenomeni ammonì che nel sottosuolo accadeva qualcosa d'importante. Il monito permise al governatore Ray e all'US Forest Service di mantenere chiusa la zona di rischio, nonostante le insistenti domande per un libero accesso. La decisione salvò sicuramente migliaia di vite.

Nei mesi successivi al 18 maggio 1980, si sono avuti ulteriori segni premonitori delle minori, ma tuttavia significative eruzioni del 12 giugno, del 22 luglio, del 7 agosto e del 16 ottobre. Il tremore vulcanico aveva preceduto di parecchie ore le eruzioni del 12 giugno e del 7 agosto, così come l'insolito incremento di piccoli terremoti proprio sotto la montagna aveva preceduto di parecchie ore le eruzioni del 22 luglio e del 16 ottobre. Anomalie nell'emissione dei gas e piccole deformazioni superficiali si erano verificate ore e giorni prima di talune delle eruzioni successive a quella del 18 maggio.

Ogni variazione nel complesso dei fenomeni premonitori deve indurre al dubbio: poi l'esperienza ci aiuterà nella valutazione. Non sono mancati, infatti, falsi allarmi, ma ciò appare inevitabile in qualsiasi sistema di previsione probabilistica.

Rappresentazione dei sei tipi di energia liberata nel corso delle manifestazioni eruttive (le date sono indicate a sinistra). L'energia sismica, l'energia deformazionale e l'energia termica delle esplosioni di vapore erano state emesse nel periodo antecedente all'eruzione del 18 maggio. La scala dell'energia è logaritmica. La riga in basso dà il totale di tutte le energie liberate dall'inizio dell'eruzione sino alla fine dell'ottobre 1980.

La sostanza fondamentale della cellula vivente

L'elevata risoluzione, raggiunta dalla microscopia elettronica ad alta tensione, rivela l'esistenza di un reticolo microtrabecolare: un sistema di filamenti leggeri e sottili che sostiene e fa muovere gli organelli

di Keith R. Porter e Jonathan B. Tucker

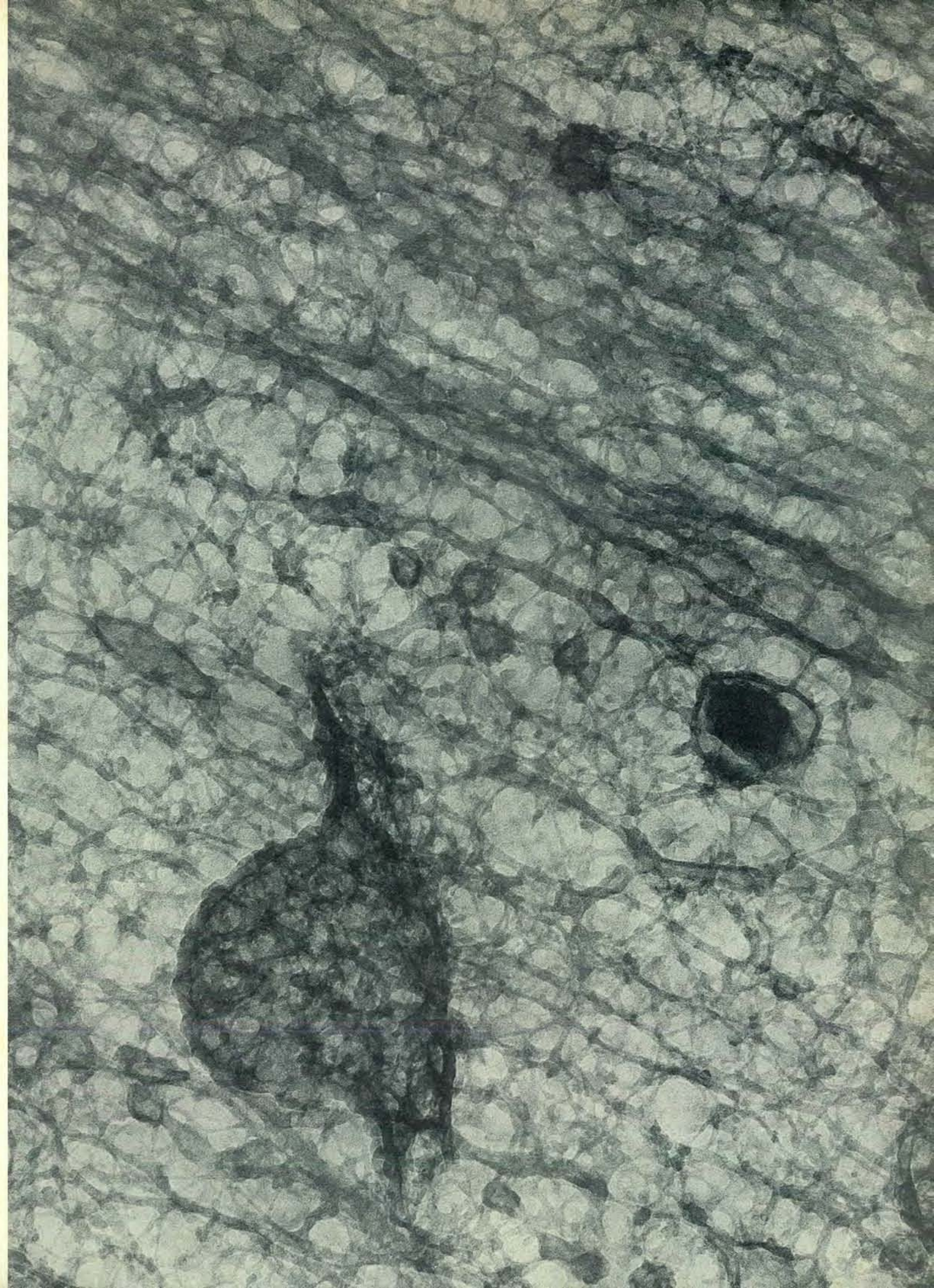
Mezzo secolo fa, la cellula vivente veniva considerata come un minuscolo sacco ripieno di liquido con solo poche strutture nel suo interno. Negli anni quaranta, l'introduzione del microscopio elettronico nelle indagini sulla cellula dimostrò che il citoplasma cellulare era organizzato in modo più rigoroso, possedendo un assortimento di vescicole delimitate da membrane e di organelli: tra gli altri, i mitocondri (cioè quelle particelle che producono ATP, la sostanza chimica che, nella cellula, funge da moneta di scambio per l'energia), i ribosomi (particelle che coordinano la sintesi delle proteine) e il reticolo endoplasmatico e l'apparato di Golgi (sistemi di vescicole che «impacchettano» le proteine da esportare dalla cellula). Il microscopio elettronico, però, rivelava ben poco riguardo alla natura del mezzo indistinto, chiaramente privo di struttura, in cui sono sospesi i vari componenti cellulari, cioè la sostanza fondamentale del citoplasma.

La sostanza fondamentale ha, invece, parecchi caratteri che fanno immediatamente pensare a una complessa struttura interna. Per esempio, esibisce proprietà viscoelastiche più vicine a quelle di un gel che a quelle di un liquido, influisce sulla forma della cellula, rigenera i componenti che vanno perduti e incorpora il meccanismo molecolare che serve al trasporto diretto di vescicole, granuli e cromosomi. Inoltre, si nota che gli organelli non possono fluire liberamente come accadrebbe in una soluzione priva di struttura, ma conservano attorno al centro della cellula una distribuzione radiale, il che fa pensare che essi siano limitati nei movimenti da un certo tipo di impalcatura interna.

Negli anni settanta, la comprensione della struttura cellulare compì un salto quantitativo quando ci si rese conto che la cellula possedeva una rete di elementi scheletrici elaborata e dinamica. Esistono perlomeno tre sistemi citoscheletrici distinti dal punto di vista chimico: i microtubuli, i microfilamenti e i filamenti intermedi. Questi elementi strutturali sono coinvolti nella mobilità cellulare e nella capacità delle cellule di conservare la propria forma. Nella sostanza fondamentale del citoplasma non venne trovata però alcuna struttura che potesse integrare e tenere sotto controllo l'assemblaggio di queste differenti reti di fibre in modo tale che potessero essere ritenute responsabili del comportamento delle cellule quale esso appariva.

Per chiarire questo mistero fu necessario poter usufruire di uno strumento di analisi più potente del microscopio elettronico standard, e cioè del microscopio elettronico ad alta tensione. Il primo apparecchio di questo tipo fu costruito nel 1947, ma solo agli inizi degli anni sessanta Gaston-Léopold Dupouy del Laboratoire d'Optique Électronique di Toulouse e V. E. Cosslett dell'Università di Cambridge misero a punto dei modelli per la ricerca scientifica. Le versioni correnti del microscopio elettronico ad alta tensione hanno un'altezza di 9 metri, pesano più di 20 tonnellate e sono in grado di accelerare gli elettroni, attraverso una caduta di potenziale di un milione di volt, dieci volte superiore a quella dei microscopi elettronici standard. Questa differenza di potenziale conferisce agli elettroni una sufficiente energia cinetica da farli penetrare in campioni spessi o addirittura in cellule intatte, aventi uno spessore di parecchi micrometri. (Un micrometro è un millesimo di millimetro.) L'immagine che ne risulta è analoga a una lastra ai raggi X e rivela la struttura interna della cellula senza dover ricorrere, come nel caso dei microscopi a più bassa tensione, alla dissezione della cellula suddetta in fettine più sottili di 0,2 micrometri.

Il principale vantaggio della microscopia ad alta tensione consiste nel fatto che dà informazioni in profondità, offrendo una visione senza confronti dell'organizzazione cellulare tridimensionale; inoltre, la risoluzione e la profondità di campo sono eccellenti per l'intero spessore del campione. Tuttavia, malgrado questi indubbi vantaggi, i biologi hanno sfruttato con lentezza il potenziale della microscopia elettronica ad alta tensione: tecnica che viene correntemente usata, in primo luogo, in metallurgia per lo studio di sottili pellicole di metalli. Mentre in Europa e in Giappone sono disponibili per la ricerca biologica parecchi microscopi elettronici ad alta tensione, negli Stati Uniti - per lo stesso scopo - ne vengono utilizzati sol-

In questa microfotografia di una minuscola area vicino al bordo di una cellula coltivata di ratto, scattata al microscopio elettronico ad alta tensione, il reticolo microtrabecolare appare come un intreccio di filamenti interconnessi. Il reticolo è visibile solo ai fortissimi ingrandimenti di questo strumento ad alta tensione, dato che le microtrabecole, cioè i singoli filamenti, sono estremamente sottili: hanno uno spessore di circa sei nanometri, cioè di sei milionesimi di millimetro. Qui il reticolo tridimensionale tiene in sospensione un organello cellulare, il lisosoma *(il piccolo corpo scuro a destra)* e tiene unito con legami trasversali un fascio di microfilamenti *(in alto)*. Esso sostiene anche un microcollare: un prolungamento che si estende come una manopola dalla superficie cellulare *(in basso)*. I piccoli corpiccioli tondeggianti che si notano in corrispondenza di alcune intersezioni del reticolo sono ribosomi. La microfotografia, scattata da Karen L. Anderson dell'Università del Colorado a Boulder, ingrandisce le strutture di circa 100 000 diametri.

tanto tre: uno all'Università del Colorado a Boulder, un altro all'Università dello Wisconsin a Madison e il terzo nei laboratori del New York State Department of Public Health di Albany.

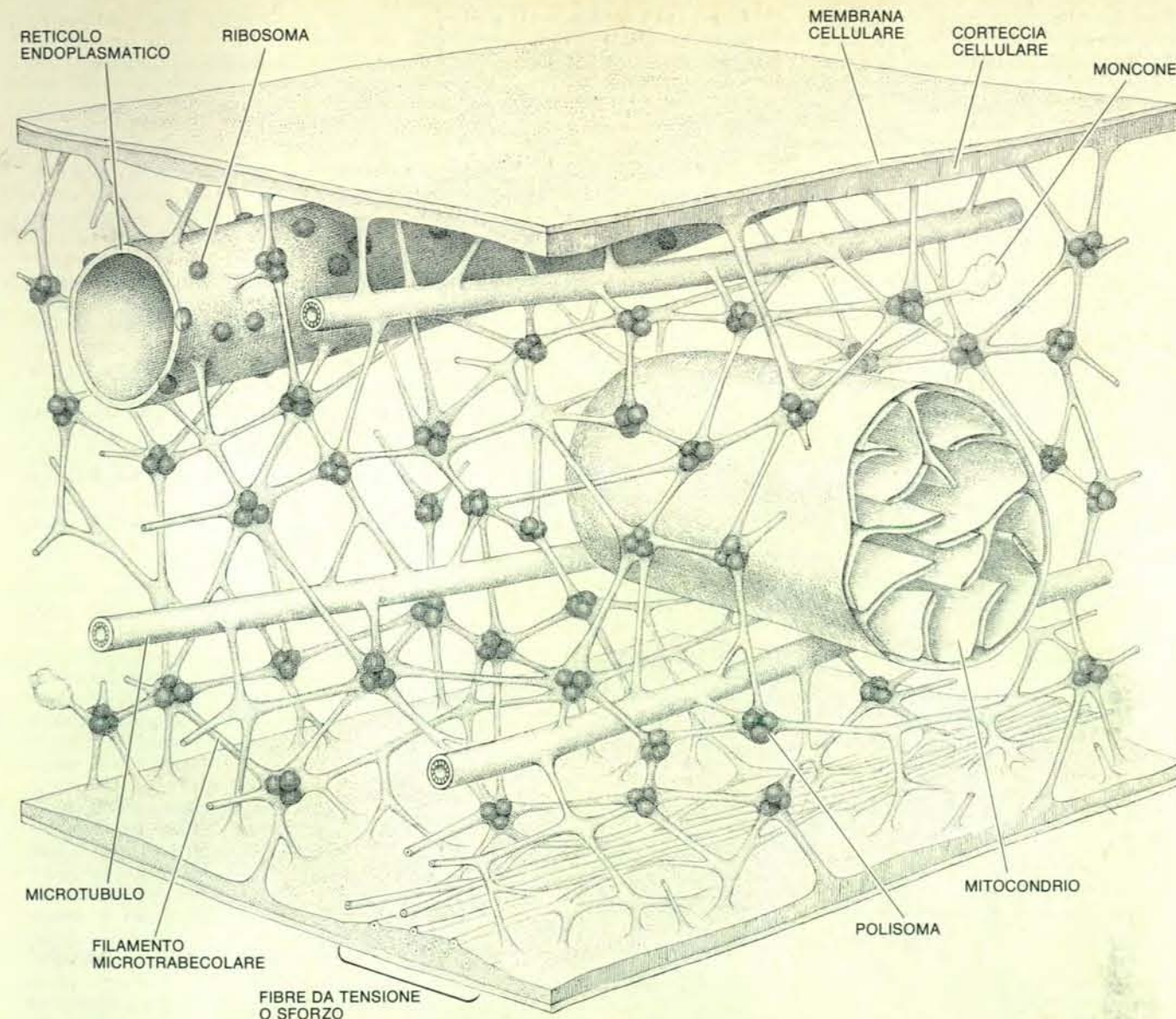
Agli inizi degli anni settanta, uno degli autori del presente articolo (Porter), lavorando con Ian K. Buckley e John J. Wolesewick, si servì appunto di uno dei suddetti microscopi, quello di Boulder, per ricercare una struttura nella sostanza fondamentale del citoplasma cellulare.

Il microscopio elettronico ad alta tensione installato a Boulder nel 1973 è uno dei tre soli microscopi di questo tipo destinati negli Stati Uniti alla ricerca biologica. Lo strumento è alto 9,6 metri e pesa 22 tonnellate. Per ridurre le vibrazioni il microscopio è dotato di fondamenta indipendenti. Gli elettroni vengono iniettati in un tubo acceleratore, che è racchiuso nel grosso serbatoio visibile al centro, che aumenta la loro energia a valori attorno a milioni di volt. Questi elettroni sono in grado di produrre immagini di sezioni spesse e addirittura di cellule intere senza perdite significative di risoluzione, cosa che non può essere fatta da elettroni con potenziali inferiori. La colonna ottica-elettronica deve essere grande per poter alloggiare le potenti lenti elettromagnetiche richieste per deflettere e per far convergere gli elettroni ad alta energia. A causa dell'intensità dei raggi X prodotta dove gli elettroni colpiscono la superficie, la colonna e il basamento del microscopio presentano un rivestimento protettivo di piombo che salvaguarda l'operatore. I caratteri giapponesi visibili al centro stanno per «montagna elettronica».

Vennero scelte per lo studio delle cellule in coltura, perché esse tendono a distendersi su una superficie piatta e pertanto possono essere osservate in tutto il loro volume. Tali cellule furono coltivate direttamente su griglie d'oro da tre millimetri, rivestite da una pellicola di plastica, fissate con glutaraldeide, colorate e ulteriormente fissate con tetrossido di osmio; quindi disidratate con alcool e acetone. Furono quindi trasferite in anidride carbonica liquida e fatte essiccare con il metodo del punto critico, con il quale si evita la distorsione dovuta alla tensione superficiale, che fa afflosciare le cellule quando vengono essiccate all'aria. Infine, le griglie furono introdotte nella camera a vuoto del microscopio, dove le cellule poterono essere osservate a grandissimo ingrandimento.

Quando i ricercatori di Boulder esaminarono piccole regioni della sostanza fondamentale del citoplasma vicino al bordo cellulare, notarono un reticolo tridimensionale irregolare di sottili filamenti che attraversavano tutto il citoplasma e si estendevano alla membrana esterna della cellula. I vari filamenti legati tra loro sembravano tener sospesi i vari sistemi e organelli cellulari. Inoltre, elementi citoscheletrici di maggiori dimensioni come i microtubuli e i microfilamenti risultavano rivestiti da una matrice estesa ai singoli filamenti del reticolo, indicando così che gli elementi più grandi costituivano un tutto unico con il reticolo ed erano essenzialmente sospesi in esso. Il reticolo faceva venire in mente la struttura trabecolare dell'osso spugnoso e venne pertanto definito reticolo microtrabecolare. Da allora la sua morfologia è stata osservata nella sostanza fondamentale di tutte le cellule degli eucarioti (cellule nucleate), prese finora in considerazione.

Quando, nel 1976, fu pubblicato su «The American Journal of Anatomy» il primo articolo che riportava la descrizione completa del reticolo microtrabecolare, numerosi ricercatori rimasero scettici. Essi sostenevano che i severi trattamenti necessari per fissare e colorare le cellule da osservare poi al microscopio avevano fatto condensare le proteine in soluzione del citoplasma attorno ai vari organelli e fibre, creando così l'erronea impressione di una matrice strutturata. I ricercatori di Boulder avviarono, di conseguenza, studi approfonditi per convalidare le loro conclusioni, ricorrendo a parecchi metodi di fissazione e di disidratazione, onde ottenere cellule intere da analizzare al microscopio elettronico ad alta tensione. In ognuno dei casi osservati, nella sostanza fondamentale appariva la morfologia del reticolo. Inoltre, le soluzioni con elevato contenuto di proteine, esposte agli stessi trattamenti, non mostrarono alcun artefatto dovuto a fissazione e simile al suddetto reticolo.

Secondo il metodo di fissazione, il reticolo varia in qualche misura, nei minimi dettagli della sua morfologia. Per esempio, nelle cellule fissate con glutaraldeide, le singole microtrabecole possono avere

Questo modello del reticolo microtrabecolare, che ha una dimensione superiore di 300 000 volte a quella effettiva, è stato ottenuto da centinaia di immagini di cellule coltivate, osservate al microscopio elettronico ad alta tensione. Esso illustra in che modo i filamenti microtrabecolari sono correlati con altri componenti del citoplasma cellulare, cioè di quella sostanza che, nella cellula, è esterna al nucleo. Le microtrabecole tengono in sospensione le strutture allungate del reticolo endoplasmatico (il sistema di canali interconnessi in cui alcune proteine sintetizzate dalla cellula stessa vengono tenute appartate), i mitocondri (organelli che sintetizzano l'ATP), i microtubuli (fibre complesse, che svolgono molte funzioni) e i microfilamenti immersi nello strato corticale della cellula (cioè lo strato di materiale immediatamente sotto la membrana esterna). In corrispondenza delle giunture del reticolo microtrabecolare si trovano i polisomi, ammassi di ribosomi, dotati di organizzazione.

uno spessore superiore ai 10 nanometri alle estremità, ma di solo due o tre nanometri nella parte centrale. (Un nanometro è un milionesimo di millimetro.) Quando le cellule vengono fissate in tempi dell'ordine dei millisecondi, congelandole in propano raffreddato alla temperatura dell'azoto liquido (–196 gradi centigradi), le microtrabecole sono più spesse e hanno un diametro più costante di circa 15 nanometri. Questa differenza nella morfologia del reticolo è chiaramente dovuta al fatto che la glutaraldeide è un fissatore piuttosto lento, che richiede alcuni secondi per esercitare in pieno il proprio effetto, mentre il congelamento è virtualmente istantaneo. Pertanto durante la fissazione con la glutaraldeide, le microtrabecole hanno il tempo di contrarsi o di allungarsi, alterando il proprio diametro e il loro rapporto reciproco e quello con gli organelli citoplasmatici.

Se il reticolo microtrabecolare è reale, perché non era stato notato nelle microfotografie elettroniche convenzionali? In realtà ciò era avvenuto, ma troppo indistintamente per poter essere riconosciuto come tale. Nelle singole sezioni sottili, le microtrabecole appaiono come strutture amorfe, a mazzi, facilmente confondibili con artefatti. L'organizzazione dei filamenti in un reticolo non appare affatto. Una parte del problema consiste nel fatto che la sezione è tanto sottile che impedisce di vedere chiaramente le interconnessioni degli elementi del reticolo. Inoltre, la resina epossidica, usata per l'inclusione nel caso delle dissezioni sottili, ha l'effetto di mascherare le strutture cellulari che hanno una capacità di dispersione degli elettroni simile a quella della resina stessa.

Nel tentativo di risolvere il problema, il gruppo di Boulder preparò sezioni sottili di cellule, includendo queste ultime in una matrice idrosolubile: il glicole polietilenico. Questo materiale, una volta ottenute le sezioni per l'osservazione al microscopio elettronico, può essere disciolto e le sezioni si possono far essiccare con il metodo del punto critico, onde evitare il loro afflosciamento. Osservate in un microscopio elettronico standard, le fettine mostrano di possedere una morfologia

reticolare, che assomiglia a quella notata in cellule intere, non incluse, analizzate al microscopio elettronico ad alta tensione. Altre tecniche di preparazione, meno distruttive, come il metodo del criodecapaggio, usato da John E. Heuser dell'Università di Washington, mettono in evidenza una struttura reticolare analoga. Così, anche se alcuni aspetti di tale struttura sono risultati degli artefatti (ed è necessario dire che i rigorosi metodi di preparazione della microscopia elettronica fanno sì che ogni cosa vista al microscopio elettronico sia in qualche misura un artefatto), restano sufficienti prove per dire che, nelle cellule, vi è in effetti una sostanza fondamentale strutturata.

Forse la prova più convincente a sostegno dell'esistenza di un reticolo microtrabecolare è il fatto che la sua struttura varia in risposta ai cambiamenti di forma della cellula o dell'ambiente cellulare. Per esempio, a bassa temperatura (4 gradi centigradi) le cellule in coltura diventano sferiche, cioè assumono la forma a più bassa energia. Un simile cambiamento si affianca allo smantellamento sequenziale delle strutture scheletriche interne della cellula. In primo luogo si scompongono i microtubuli, seguiti dai microfilamenti. Infine, il reticolo microtrabecolare si deforma, ma senza decomporsi completamente. Alcune microtrabecole interrompono le loro connessioni con il reticolo e si riuniscono in minuscole masserelle, creando così grosse lacune nel reticolo. Queste rendono possibile il moto browniano (diffusione libera) degli organelli citoplasmici, un fenomeno che si riscontra nelle cellule raffreddate, ma non in quelle a temperature fisiologiche. Se una cellula viene riportata per soli cinque minuti alla temperatura del corpo umano (37 gradi centigradi), si osserva una ristrutturazione improvvisa e marcata del reticolo, che torna alla sua morfologia iniziale.

Analogamente, effetti marcati sulla struttura del reticolo microtrabecolare possono essere sollecitati dalla citocalasina B, un farmaco che inibisce un'ampia varietà di movimenti cellulari. Dopo 10 minuti di esposizione alla sostanza, il reticolo assume un aspetto molto grossolano: le microtrabecole si ispessiscono e si allungano e, di conseguenza, gli spazi intertrabecolari si allargano. Subito dopo la rimozione della citocalasina, il reticolo torna alla sua struttura normale finemente suddivisa. Altri agenti e condizioni (pressione osmotica eccessivamente alta o bassa, livelli alterati a proposito di ioni specifici come il magnesio e il calcio e presenza di inibitori metabolici) inducono anch'essi nel reticolo modificazioni strutturali uniche e reversibili. Se il reticolo fosse semplicemente un artefatto dovuto alla fissazione, sarebbe difficile trovare una spiegazione a queste ampie fluttuazioni della sua morfologia.

Quali sono le implicazioni, per la struttura e la funzione della cellula, del reticolo microtrabecolare? Sembra evidente che esso suddivide la cellula in due fasi: la fase del reticolo stesso, ad alto contenuto proteico e polimerizzata, e la fase liquida, ad alto contenuto d'acqua, che riempie gli interstizi del reticolo (spazi intertrabecolari). Con un contenuto d'acqua del 50 per cento, questa struttura conferisce al citoplasma la sua tipica consistenza di gel.

La realtà di queste due fasi venne dimostrata dal gruppo di Boulder (ivi compresa Karen L. Anderson) in esperimenti in cui il contenuto idrico delle cellule veniva modificato. Le cellule venivano immerse in soluzioni contenenti concentrazioni basse o elevate di molecole di zucchero, il che faceva entrare o uscire acqua per osmosi. Esse venivano poi esaminate con il microscopio elettronico ad alta tensione, il che permetteva di notare che gli spazi intertrabecolari si espandevano quando l'acqua penetrava all'interno della cellula e si restringevano quando, invece, l'acqua veniva risucchiata verso l'esterno. Quando l'acqua veniva rimossa dalle cellule e rimpiazzata subito dopo, le cellule riuscivano a sopravvivere: osservazione che può trovare una spiegazione nella capacità del reticolo di preservare la vitalità cellulare proteggendo la cellula dalle normali fluttuazioni di contenuto idrico.

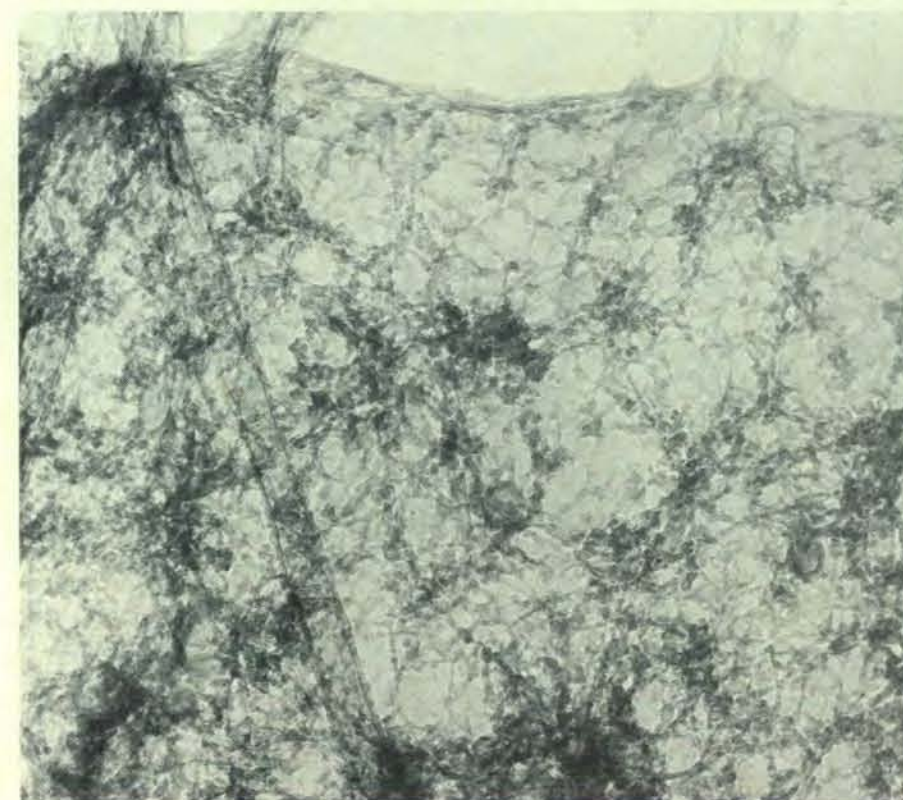
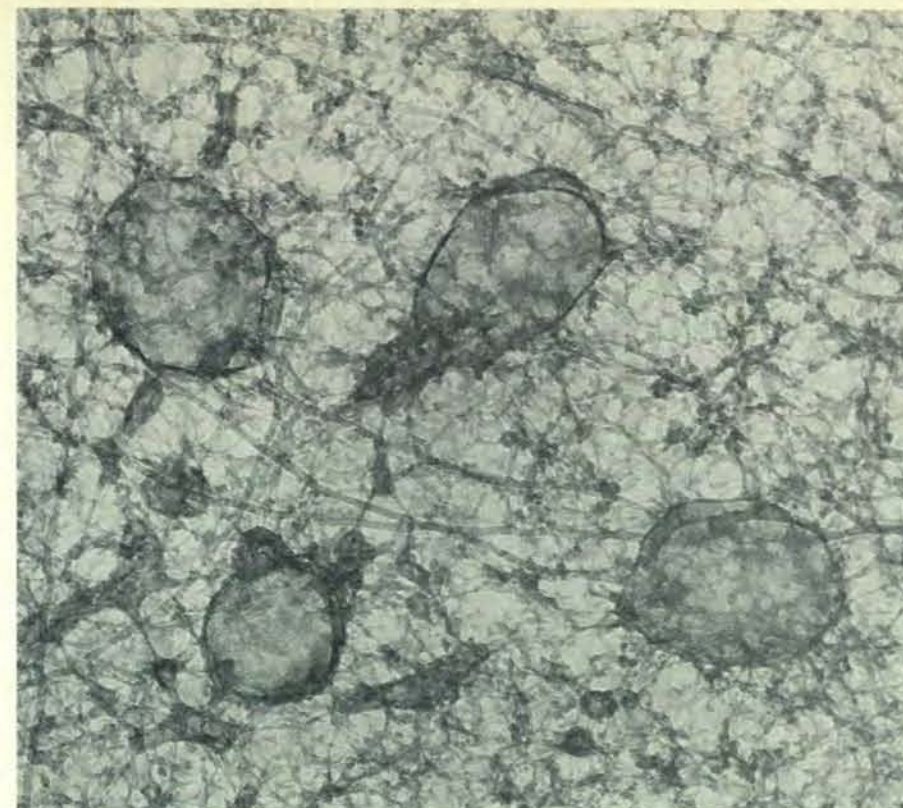
Il reticolo microtrabecolare può anche avere un ruolo nella organizzazione degli enzimi in seno al citoplasma. Nel 1929, Rudolph Peters dell'Università di Cambridge sosteneva su basi puramente logiche che le intricate reazioni biochimiche che - come si sa - si svolgono nel citoplasma avevano un'organizzazione troppo rigida per poter essere governate da collisioni casuali di enzimi e dei loro substrati (cioè delle sostanze di cui effettuano la trasformazione), che si diffondono nel mezzo astrutturato. Egli postulava l'esistenza di «una certa tenue rete, la cui azione coordina le varie attività enzimatiche della cellula». Le congetture di Peters erano decisamente profetiche: si hanno oggi valide prove che un certo numero di enzimi interessati nelle vie metaboliche (per esempio la via della glicolisi, cioè della conversione del glucosio in piruvato) è legato a strutture che sono interne alla sostanza fondamentale del citoplasma. Ne consegue che tali enzimi possono essere incorporati nel reticolo microtrabecolare, forse con un orientamento non casuale, per cui un enzima passa il proprio substrato all'enzima successivo della via metabolica. In questo modo, gli enzimi che non risultano legati alle membrane cellulari potrebbero nondimento essere coordinati da un punto di vista spaziale.

Microfotografia al microscopio elettronico convenzionale di una sottile sezione effettuata nello spessore della sostanza fondamentale del citoplasma di una cellula umana in coltura: il reticolo microtrabecolare risulta poco evidente anche se si possono notare filamenti più spessi assieme a parti del reticolo endoplasmatico e a mitocondri. La sostanza fondamentale mostra numerosi filamenti a ciuffi, distribuiti casualmente, alcuni dei quali decorrono tra i microtubuli *(strutture a bastoncello)*. La resina epossidica nella quale le strutture sono incluse per l'osservazione maschera le piccole aree. La microfotografia, di John J. Wolosewick, dà un ingrandimento di 14 000 diametri.

La fase acquosa che riempie gli spazi intertrabecolari contiene piccole molecole come il glucosio, gli amminoacidi, l'anidride carbonica e l'ossigeno. A causa della struttura del reticolo, le distanze che le molecole percorrono per diffondersi in tale fase sono brevi. Ciò ha come effetto un'accelerazione della velocità con la quale i substrati passano da un enzima al successivo.

Il reticolo sembra anche svolgere un importante ruolo nel differenziamento cellulare e nella sintesi proteica a esso associata: cioè nel processo con il quale i polisomi (grappoli di ribosomi legati a un filamento di RNA messaggero) traducono il codice genetico nella sequenza specifica di amminoacidi che costituisce una molecola proteica. Sulla base di microfotografie elettroniche di sezioni sottili, si è creduto in generale che il processo di traduzione fosse mediato da polisomi liberamente sospesi nel citoplasma. Microfotografie scattate al microscopio elettronico ad alta tensione rivelano una morfologia sorprendentemente diversa: i polisomi giacciono in corrispondenza delle intersezioni del reticolo microtrabecolare, come ragni in una ragnatela.

Questa notevole organizzazione presenta delle implicazioni significative, dato che le proteine, una volta che sono state sintetizzate, entrano a far parte del materiale associato con il reticolo microtrabecolare, il quale diventa allora disponibile in modo non casuale per l'assemblaggio controllato di microtubuli, microfilamenti e altre strutture cellulari. Per esempio, i microfilamenti si possono organizzare in fasci paralleli, chiamate fibre da tensione o sforzo, oppure in una rete amorfa. A mano a mano che la cellula passa ad altre attività, vengono continuamente riuniti e disuniti in parti diverse del citoplasma. L'esame della funzione del reticolo microtrabecolare nell'assemblaggio delle strutture cellulari può avere un'importanza cruciale nella comprensione del modo in cui le cellule si accrescono e si specializzano.

Dato che il reticolo occupa l'intera cellula, sembra legare i disparati componenti del citoplasma - organelli e fibre strutturali - in un'unica unità strutturale e funzionale, detta citoplasto. Qual è il rapporto tra il reticolo e l'elaborata rete citoscheletrica, formata dai tre principali si-

L'effetto delle basse temperature sul reticolo microtrabecolare, forse come conseguenza di un arresto, indotto dal freddo, della macchina metabolica della cellula, è drammatico. Le cellule normali raffreddate a 4 gradi centigradi per tre ore assumono una forma sferica, e subiscono insieme numerose modificazioni interne. Innanzitutto i microtubuli si smontano nello spazio di alcuni minuti; compaiono quindi nel reticolo grosse lacune, come risulta dalla microfotografia (al microscopio elettronico ad alta tensione) in basso. La morfologia normale appare, invece, nella microfotografia in alto. Le microtrabecole staccate si contraggono in grossi ammassi allorché le singole trabecole si trasformano in minuscoli «monconi». Quando le cellule raffreddate vengono incubate di nuovo a 37 gradi centigradi, i monconi si ritrasformano in microtrabecole; dopo 30 minuti di incubazione il reticolo è essenzialmente normale. La sua sensibilità ai cambiamenti dell'ambiente cellulare fa pensare che sia una componente fisiologicamente attiva della cellula. Le microfotografie, scattate dalla Anderson, ingrandiscono le strutture di circa 78 000 diametri.

Gli spazi tra le microtrabecole *(in alto)* contengono una grande quantità d'acqua con molte molecole essenziali per il metabolismo cellulare. Essendo piccole le dimensioni degli spazi, la diffusione delle suddette molecole da un enzima al successivo richiede solo dei microsecondi. L'esistenza degli spazi intertrabecolari è stata dimostrata alterando la concentrazione del mezzo extracellulare e facendo così entrare e uscire acqua dalla cellula per osmosi. Se la concentrazione all'esterno della cellula viene aumentata a un livello superiore a quello del citoplasma, l'acqua viene risucchiata fuori dalla cellula e gli spazi intertrabecolari si riducono *(al centro)*. Se, al contrario, la concentrazione diminuisce l'acqua penetra nella cellula e gli spazi si dilatano *(in basso)*. L'ingrandimento delle microfotografie, scattate dalla Anderson, è di circa 35 000 diametri.

stemi di fibre? Di recente, Manfred Schliwa, lavorando a Boulder, ha preso in esame questo problema, sottoponendo cellule in coltura a estrazione con un detergente (Triton X-100) in presenza di un tampone che rende stabili i componenti citoscheletrici. Il trattamento con quel detergente per 20 secondi ha rimosso virtualmente dalle cellule tutte le strutture delimitate da membrane, ivi compresa la membrana esterna, i mitocondri, il reticolo endoplasmatico e l'involucro del nucleo cellulare. Sono stati estratti inoltre quasi tutti i filamenti microtrabecolari, il che fa pensare che essi abbiano proprietà fisiche e una composizione chimica ben diverse da quelle degli elementi citoscheletrici stabili.

Il materiale non estratto costituiva un'intelaiatura filamentosa, che si avvicina alla conformazione della cellula ed era costituita da microfilamenti, microtubuli e filamenti intermedi. Al microscopio elettronico ad alta tensione l'organizzazione tridimensionale dei suddetti filamenti appariva ben conservata e virtualmente identica a quella della cellula intatta. Inoltre, i vari filamenti citoscheletrici risultavano interconnessi in corrispondenza di molteplici siti, spiegando così perché la rottura di un unico sistema di filamenti con inibitori specifici può indurre cambiamenti nell'organizzazione anche di un altro sistema di filamenti.

I filamenti microtrabecolari, assieme agli elementi citoscheletrici, sembrano avere una funzione di sostegno, con adattamento alla forma della cellula, per la membrana cellulare esterna e per i prolungamenti, marcatamente asimmetrici, della superficie cellulare. Per esempio, le ciglia e i flagelli, appendici a forma di frusta che fungono da organelli di propulsione per i protozoi e gli spermatozoi, sono strutture complesse, costituite da molteplici serie di microtubuli, legate da «raggi» e da «ponti». Quando queste strutture specializzate vengono sezionate assieme a una matrice di glicole polietilenico ed esaminate al microscopio elettronico ad alta tensione, raggi e ponti appaiono esattamente come microtrabecole. Analogamente, i fasci di microfilamenti nei microvilli, prolungamenti citoplasmatici che si osservano nelle cellule intestinali, sembrano associati alle microtrabecole. Tutte osservazioni, queste, che fanno pensare che la morfologia interna delle ciglia, dei flagelli, dei microvilli e di altri prolungamenti cellulari specializzati sia identica a quella del citoplasto come insieme: consiste di microtrabecole e di spazi intertrabecolari.

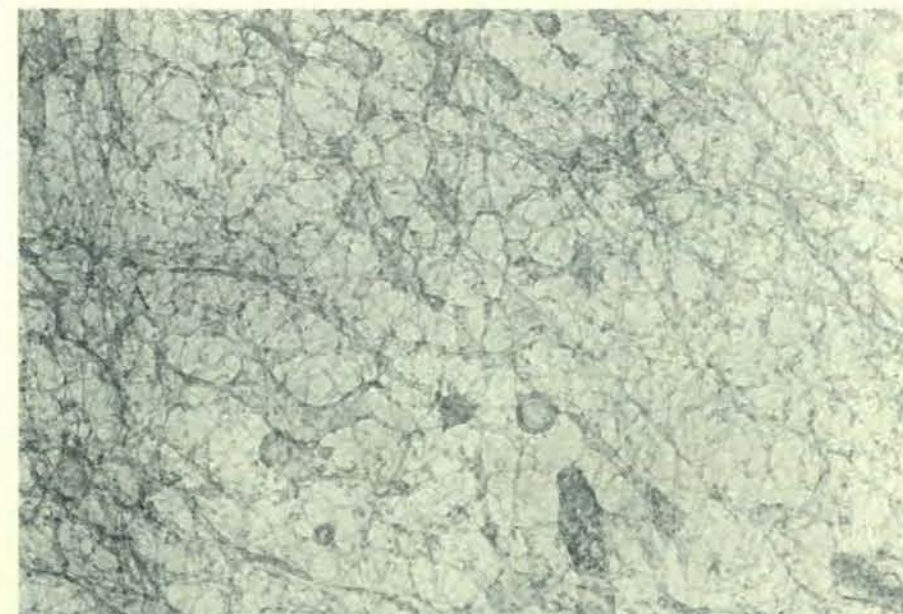
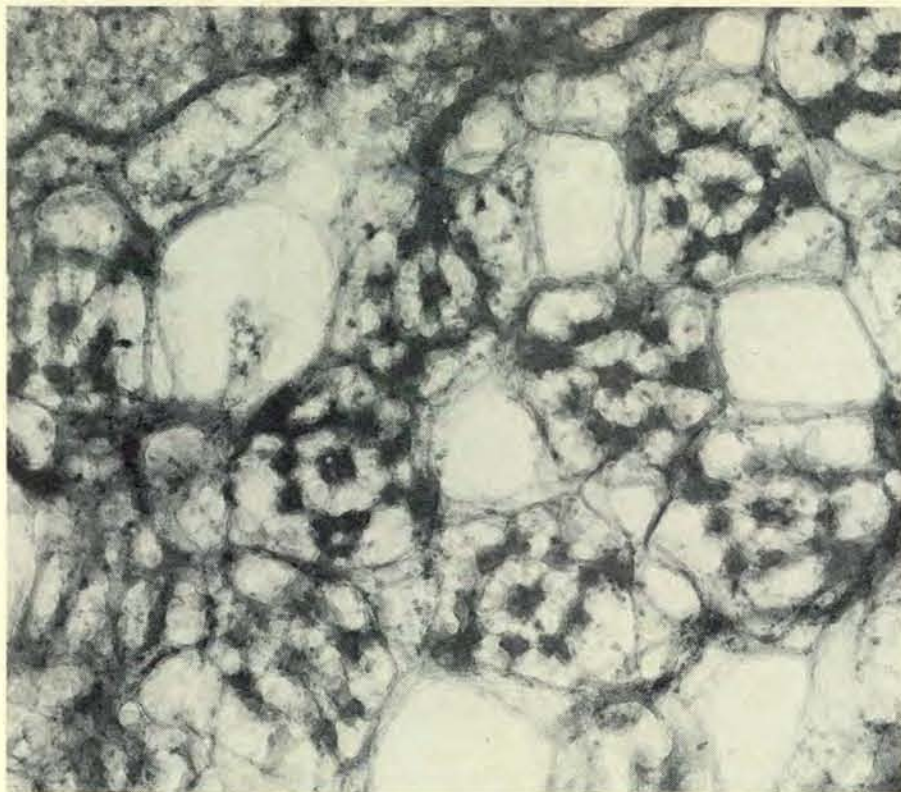
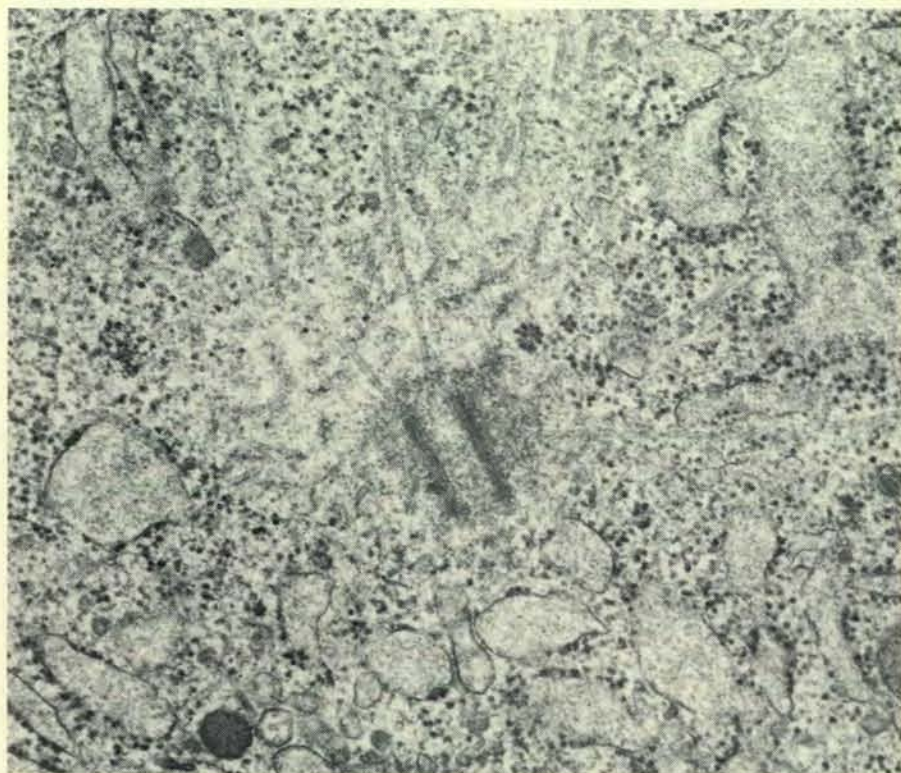
Un organello citoplasmatico che sembra intimamente associato con il reticolo microtrabecolare è il centro organizzatore del microtubulo (COMT), vicino al centro della cellula. Un tipico COMT, quale lo si osserva in cellule animali, consta di una coppia di centrioli (ciascuno consistente di una serie di microtubuli, disposti a costituire un cilindro), orientati ad angolo retto l'uno rispetto all'altro e circondati da un complesso di dense proteine satelliti, che variano di forma, numero e disposizione da un tipo cellulare all'altro. I centrioli e i loro satelliti fungono da siti di inizio per l'assemblaggio dei microtubuli, ed è la distribuzione di questi elementi citoplasmatici secondo determinati modelli a influenzare la conformazione della cellula.

I COMT hanno una continuità con il reticolo microtrabecolare: non si ha tra loro alcuna demarcazione strutturale netta. Inoltre, l'informazione per la localizzazione di questo complesso al centro della cellula sembra essere incorporata nella sostanza fondamentale. Nella porzione di citoplasma che circonda immediatamente il COMT, il reticolo microtrabecolare appare estremamente denso. Come risultato, questa parte della cellula, nota come centrosfera, è gelatinosa. La densità del reticolo diminuisce, invece, con l'aumentare della distanza dal COMT verso la periferia.

Nei primi stadi della duplicazione cellulare, la centrosfera si divide. La sua duplice progenie forma i poli del fuso mitotico, che dirigono la separazione dei cromosomi, anch'essi duplicati, nel corso della mitosi o divisione cellulare. Al concludersi di questa, ogni cellula figlia contiene una centrosfera circondata dal proprio reticolo microtrabecolare. Questa osservazione fa sorgere un interrogativo provocatorio: qual è l'origine dell'informazione strutturale immagazzinata nel reticolo? Viene codificata nei geni o viene trasmessa fisicamente da una generazione alla successiva nel corso della divisione cellulare che si svolge durante lo sviluppo embrionale? Va notato che alcuni esperimenti fanno pensare piuttosto a quest'ultima possibilità.

Nel 1964, T. M. Sonneborn e collaboratori dell'Università dell'Indiana effettuarono alcuni eleganti esperimenti sulla trasmissione ereditaria di strutture cellulari nel protozoo ciliato *Paramecium*. Questo animale unicellulare è complesso e possiede tra le altre cose numerose ciglia attaccate allo strato corticale della cellula, cioè a quella parte del reticolo microtrabecolare che si trova immediatamente sotto la membrana cellulare esterna. Sonneborn e collaboratori hanno eseguito degli interventi microchirurgici sulle cellule di paramecio, alterando in vario modo il normale tipo di struttura per ottenere, ad esempio, cellule con due o più aperture boccali o anali, o sezioni di strato corticale con ciglia dotate di polarità invertita. Essi pensavano che questi sconvolgimenti imposti nella normale struttura della cellula sarebbero stati letali o rapidamente corretti dai geni cellulari. Con loro grande sorpresa, invece, le anomalie indotte venivano ereditate dalla progenie in occasione di divisioni cellulari successive! I cambiamenti si perpetuavano così, attraverso le divisioni, per più di un anno, cioè per più di 700 generazioni di cellule.

Sonneborn concluse che l'informazione sulla forma viene trasmessa da una generazione alla successiva attraverso un meccanismo non genetico: l'informazione di base sulla forma della cellula è immagazzinata nella struttura della sostanza fondamentale del citoplasma, la quale - in corrispondenza della mitosi - viene trasmessa a tutte e due le cellule figlie. Questo ordinamento e questa disposizione della nuova struttura cellulare sotto l'influenza di una struttura cellulare preesistente sono state definite citotassia. Sonneborn ha scritto: «Mentre i geni determinano le unità costruttive molecolari e, attraverso le loro proprietà, i tipi di associazioni molecolari che possono verificarsi, le associazioni che, nella realtà, si costituiscono dipendono anche da quelle già esistenti».

L'idea che la struttura e la motilità citoplasmatiche fossero correlate tra loro risale a uno dei primi citologi francesi, Félix Dujardin, il quale nel 1835 avanzò la proposta che tutte le cellule fossero costituite da un materiale, che chiamò *sarcode*, dotato di proprietà sia strutturali sia contrattili. Risulta oggi che il reticolo microtrabecolare svolge un simile ruolo,

Il trattamento con un detergente mette in luce il rapporto esistente tra le microtrabecole e i filamenti del citoscheletro: microtubuli, microfilamenti e filamenti intermedi. La microfotografia in alto mostra la cellula non trattata con il reticolo microtrabecolare e gli organelli intatti. Quella in basso mostra, invece, una cellula analoga, ma trattata con il detergente Triton X-100, che allontana il materiale del reticolo che ricopre i filamenti citoscheletrici, gli organelli e le membrane. Rimangono gli elementi citoscheletrici che costituiscono una rete, la quale ha una forma che si avvicina a quella della cellula. I microtubuli, i microfilamenti e i filamenti intermedi si possono distinguere in funzione del loro differente spessore. Le microfotografie, scattate da Manfred Schliwa all'Università del Colorado a Boulder, ingrandiscono le strutture di circa 40 000 diametri.

I flagelli degli spermatozoi di topo, la cui estremità sono visibili dall'alto in questa microfotografia ottenuta con un microscopio elettronico ad alta tensione, constano di serie di microtubuli, collegati da strutture simili a raggi, che assomigliano molto a microtrabecole. Questo fatto indica che i prolungamenti specializzati del citoplasma, quali sono i flagelli, le ciglia e i microvilli, hanno tutti una struttura reticolare, che penetra nel resto della cellula. La microfotografia che ingrandisce le strutture di circa 55 000 diametri, è stata scattata da John J. Wolosewick a Boulder.

Il centriolo, che incorpora le due strutture parallele, è visibile in corrispondenza del centro *(un poco più in basso a destra)* in questa microfotografia elettronica convenzionale di una sezione sottile di cellula. Fa parte di una coppia di centrioli, orientati a formare un angolo retto: con il materiale «satellite», formano un complesso che dirige l'assemblaggio e l'orientamento dei microtubuli. Tre microtubuli emergono dal materiale satellite che circonda le strutture parallele del centriolo. In vicinanza di tale complesso, il reticolo è denso. L'ingrandimento è di circa 36 000 diametri.

funzionando come un tipo di muscolatura cellulare dinamica che ridistribuisce e riorienta in continuazione gli organelli e le fibre citoscheletriche a mano a mano che la cellula procede nelle sue diverse attività. In nessun senso il reticolo è statico, ma piuttosto si presume che sia in costante mutamento, esibendo contrazioni e deformazioni locali, con la conseguenza di un aumento o di una diminuzione delle dimensioni degli spazi intertrabecolari.

Le cellule sono in grado di svolgere movimenti interni in molteplici modi, da quelli lenti, continui dei cromosomi durante la mitosi a quelli discontinui, che fanno pensare a un inseguimento su tracce, delle vescicole e dei granuli lungo le fibre nervose. I cromatofori, o cellule pigmentate, interessano in modo particolare chi studia la biologia cellulare perché hanno movimenti interni sia continui che improvvisi. La conversione dell'energia chimica ed elettrica in lavoro meccanico, che consiste nel trasportare i granuli di pigmento all'interno dei cromatofori, è dunque un importante tema di ricerca.

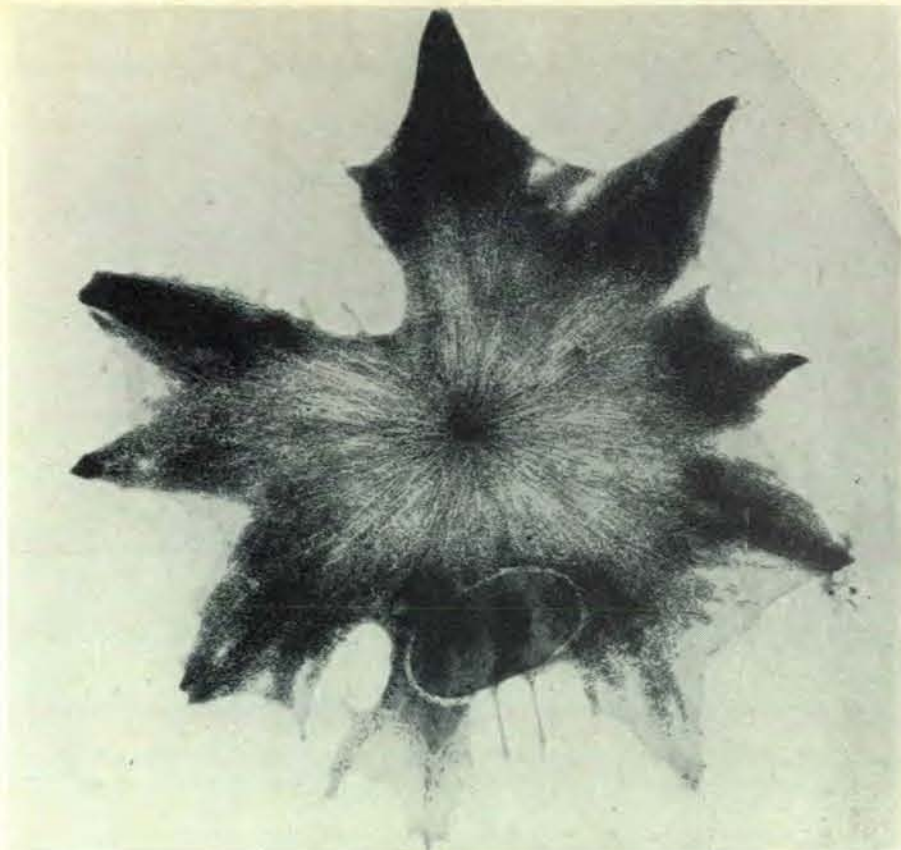
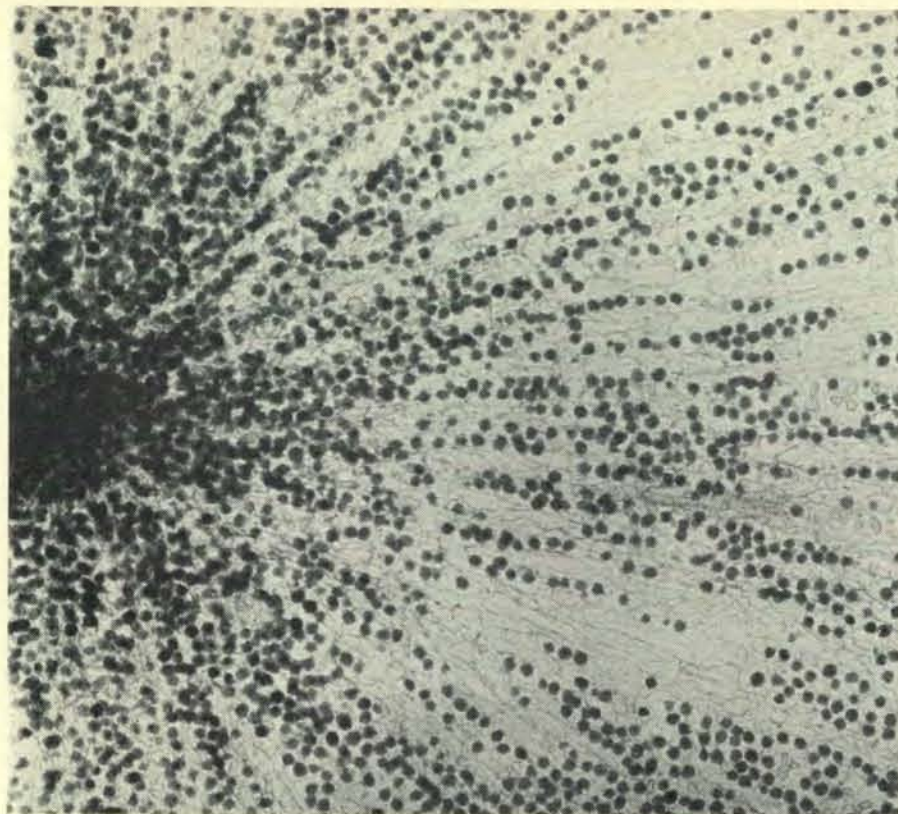
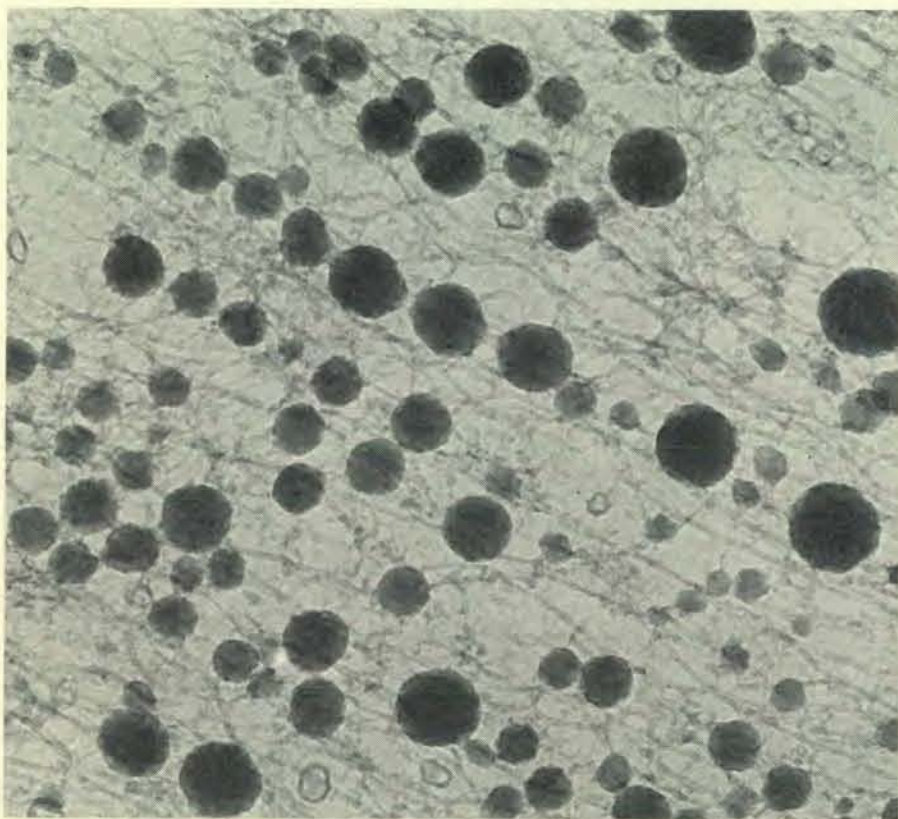
I cromatofori, presenti nella cute di certi anfibi e nelle scaglie di molti pesci, come pure nelle squame dei rettili, agiscono sotto controllo ormonale e nervoso e fanno intensificare o schiarire il colore della cute di questi animali, in modo da potersi confondere con il colore dell'ambiente e da riuscire a sfuggire meglio di fronte ai predatori. In esperimenti condotti dal gruppo di Boulder vennero scelti gli eritrofori, cellule pigmentate rosse, del pesce *Holocentrus ascensionis,* perché si isolano facilmente dall'animale e si possono conservare a lungo in coltura. Come risposta all'ormone adrenalina, migliaia di granuli di un rosso vivo, dispersi nel citoplasma dell'eritroforo, si aggregano al centro della cellula. Ognuno accelera i propri movimenti fino a raggiungere una velocità tra 15 e 20 micrometri al secondo e percorre l'intera distanza in linea retta e senza arrestarsi.

La velocità costante dei granuli che si aggregano è analoga al movimento dei cromosomi nella divisione cellulare, ma è 500 volte più rapida. In 3-5 secondi l'aggregazione è completa: i granuli si riuniscono a grappolo in una massa protuberante al centro della cellula, circondata da citoplasma chiaro; la cellula conserva il suo contorno originario, quasi circolare. Il nucleo cellulare e i mitocondri non si spostano con i granuli: o vengono lasciati nel citoplasma periferico o si muovono in maniera indipendente verso il centro della cellula. Quando l'aggregazione si verifica in maniera simultanea in milioni di eritrofori, il pesce può modificare il proprio colore dal rosso vivo al bianco assoluto, nello spazio di pochi secondi. La dispersione del pigmento avviene invece in maniera più lenta dell'aggregazione: essa richiede dai sei ai dieci secondi per completarsi e ogni granulo si sposta in maniera discontinua, con molte partenze e molti arresti e perfino, su brevi distanze, con movimenti che fanno pensare a inseguimenti.

Gli eritrofori aggregano il loro pigmento in condizioni di bassa energia cellulare, come quando la cellula muore o quando è stata sottoposta a una prolungata esposizione a un inibitore metabolico. Pertanto il processo di aggregazione è alimentato dall'energia potenziale immagazzinata, che viene messa in libertà. Per contro, la dispersione consuma ATP e immagazzina energia potenziale. La sua natura lenta e discontinua può derivare dal suo dover dipendere dall'energia.

Dalle microfotografie elettroniche standard risulta chiaramente che l'eritroforo contiene un sistema radiale di microtubuli, che si diparte dal centro della cellula, in modo simile ai raggi solari. Basandosi su questa morfologia, alla fine degli anni sessanta alcuni ricercatori hanno avanzato l'ipotesi che l'aggregazione del pigmento fosse generata da un certo tipo di interazione tra granuli di pigmento e microtubuli il che faceva pensare a uno scorrimento. Una simile ipotesi viene oggi considerata improbabile perché esistono, in realtà, scarsi contatti effettivi tra microtubuli e granuli, seppure ve ne sono.

Nel 1977, nel tentativo di indagare sulla possibile connessione tra reticolo microtrabecolare e migrazione del pigmento nell'eritroforo, il gruppo di Boulder (con H. Randolph Byers) ha esaminato al microscopio elettronico ad alta tensione degli eritrofori isolati intatti. Come in altre cellule, hanno trovato una rete di filamenti microtrabecolari sottili e allungati, che riempie gli spazi liberi e che connette tra loro i granuli di pigmento, i microtubuli e le membrane cellulari. Per approfondire lo studio del meccanismo di migrazione del pigmento, i ricercatori di Boulder hanno fissato gli eritrofori in vari stadi di aggregazione e di dispersione del pigmento stesso, esaminandoli poi al microscopio elettronico ad alta tensione: ne è emerso che il movimento dei granuli è accompagnato da profondi cambiamenti strutturali del reticolo.

Durante l'aggregazione, le microtrabecole si accorciano e si ispessiscono, trascinando i vari granuli in modo che si riuniscano insieme in grappoli e trasportandoli passivamente verso il centro della cellula lungo percorsi lineari, definiti dalla disposizione radiale dei microtubuli. A mano a mano che il reticolo si contrae, numerose microtrabecole interrompono le loro connessioni con i granuli e si ritraggono alla superficie dei microtubuli e dello strato corticale della cellula, formando «monconi» del diametro di 20 nanometri. Dopo che i granuli e il materiale del reticolo a essi associato si sono ritirati dalla periferia cellulare, lo strato corticale superiore della cellula si affloscia su quello inferiore e, quando ciò accade, i microtubuli disposti radialmente, che non si spostano con i granuli, diventano subito evidenti. D'altra parte, il reticolo in questo momento si nota solo in quelle regioni della cellula che sono immediatamente adiacenti alla massa centrale dei granuli e lungo alcuni fasci di microtubuli. In quelle aree, il reticolo

Il movimento dei granuli di pigmento dell'eritroforo, o cellula pigmentata di rosso, del pesce *Holocentrus ascensionis* viene illustrato da queste due microfotografie, scattate al microscopio elettronico ad alta tensione e relative a due cellule intere. Nella cellula in alto i granuli di pigmento sono allineati lungo le file radiali di microtubuli. Il nucleo cellulare *(corpo a forma di losanga)* si situa verso la periferia della cellula e non sembra interferire con l'allineamento dei granuli. Nella cellula della microfotografia in basso, invece, i granuli sono tutti ammassati al centro della cellula, a formare un insieme compatto. Le microfotografie, che ingrandiscono entrambe le cellule di circa 3500 diametri, sono state scattate da Katherine J. Luby-Phelps a Boulder.

La disposizione lineare dei granuli di pigmento nell'eritroforo è evidente in questa microfotografia al microscopio elettronico ad alta tensione, che li mostra allo stato disperso. L'ammasso di granuli a sinistra corrisponde al centro della cellula. Fasci di microtubuli sporgono verso la periferia e il reticolo microtrabecolare della sostanza fondamentale è evidente tra granuli e microtubuli. La microfotografia (ingrandimento circa 17 500 diametri) è opera della Luby-Phelps.

In questa microfotografia, scattata dalla Luby-Phelps al microscopio elettronico ad alta tensione, i granuli sferici di pigmento sono inclusi nel reticolo microtrabecolare. Essi appaiono rivestiti dal materiale del reticolo e alcuni sono chiaramente tenuti uniti dalle microtrabecole. I microtubuli decorrono in diagonale, da sinistra in alto a destra in basso. L'ingrandimento è di 70 000 diametri.

appare estremamente compatto: le microtrabecole hanno una lunghezza pari a circa la metà della loro lunghezza normale e, a intervalli regolari, presentano «monconi» di 20 nanometri di diametro.

Allorché la dispersione ha inizio, le singole microtrabecole si allungano, riportando di nuovo i granuli di pigmento verso la periferia della cellula e ricostruendo il reticolo. Nel contempo, vi è un allungamento dei monconi da 20 nanometri, che erano rimasti in periferia. Assieme a essi, la disposizione radiale dei microtubuli funge chiaramente da «memoria» strutturale per guidare la ricostruzione del reticolo. In effetti, quando la popolazione dei microtubuli viene smembrata dalla colchicina, i granuli risultano disorganizzati e non riescono a disperdersi.

Le ricerche svolte da Katherine J. Luby-Phelps a Boulder hanno rafforzato l'ipotesi che il reticolo agisca da mediatore nella migrazione del pigmento. Sono stati studiati gli effetti, sulla migrazione del pigmento, di inibitori metabolici usati per esaurire le scorte di ATP dell'eritroforo. Le cellule trattate mostrano un reticolo con un aspetto slegato, caratterizzato da una notevole percentuale di microtrabecole incomplete o con estremità libere: pertanto esse possono disperdere il loro pigmento solo parzialmente. In mancanza di ATP, i filamenti microtrabecolari singoli sono chiaramente incapaci di allungare e ristrutturare il reticolo. Pertanto sembra che, a spese dell'energia chimica (ATP), nel reticolo durante la dispersione venga accumulata energia potenziale che successivamente viene trasformata in energia cinetica durante l'aggregazione. Questi movimenti avanti e indietro dei granuli di pigmento nell'eritroforo a riposo possono essere il risultato di un «tiro alla fune» continuo tra la tendenza entropica del reticolo a contrarsi e la sua capacità di espandersi in risposta all'input energetico, alla stessa stregua di un laccio di gomma che viene tirato.

La contrattilità del reticolo può essere tenuta sotto controllo dagli ioni calcio: è noto, infatti, che essi regolano la contrazione muscolare e una varietà di attività motorie in cellule non muscolari. In effetti, la Luby-Phelps ha trovato che un aumento nella concentrazione degli ioni calcio, liberi nell'eritroforo, innesca la contrazione delle singole microtrabecole e porta all'aggregazione del pigmento. Il fabbisogno di ATP durante la dispersione può essere, almeno in parte, dovuto alla necessità di alimentare il meccanismo di assunzione che rimuove gli ioni calcio liberi dagli spazi intertrabecolari, in modo che le microtrabecole possano ritornare nel loro primitivo stato allungato.

La fonte più probabile di ioni calcio liberi è un sistema di canali tubulari, delimitato da una membrana che traccia il proprio cammino in mezzo ai granuli di pigmento contenuti all'interno dell'eritroforo. Un sistema poligonale di questi tubuli si estende per tutta la cellula ed è in contatto con il reticolo e, in alcuni punti, anche con lo strato corticale superiore e inferiore della cellula. Esso potrebbe fornire il calcio a tutte le parti della cellula in risposta a un impulso nervoso. (L'eritroforo possiede numerose connessioni nervose.) Sotto questo aspetto il sistema di tubuli dell'eritroforo può funzionare in maniera analoga al reticolo sarcoplasmatico del muscolo, che reagisce a un impulso nervoso emettendo calcio, il quale innesca la contrazione muscolare.

Una spiegazione completa del modo in cui il reticolo microtrabecolare si espande e si contrae attende ancora una comprensione, nei minuti particolari, della composizione molecolare e della struttura del reticolo stesso. Questa organizzazione a livello biochimico deve chiaramente permettere un comportamento elastico e, nel contempo, incoraggiare la formazione di interconnessioni multiple, sia con il reticolo sia con i microtubuli, i microfilamenti, i filamenti intermedi e gli organelli.

Quando le cellule vengono trattate con un detergente e le proteine estratte separate grazie a una elettroforesi bidimensionale su gel, si ottengono più di 100 proteine diverse. Ciò indica che il reticolo è un polimero complesso, costituito da più proteine. Fino a oggi ne è stato identificato solo un piccolo numero, ivi compresa l'actina e la miosina (le principali proteine contrattili delle cellule muscolari, che però sono universalmente diffuse anche in cellule non muscolari), la tubulina (unità costitutiva dei microtubuli) e due proteine specifiche associate con i microtubuli. Benché buona parte della struttura della matrice abbia l'aria di rappresentare una associazione non molto compatta di proteine facilmente estraibili, vi è anche la prova che, nel reticolo, esista un vero e proprio scheletro filamentoso. Molte microtrabecole possono avere una parte centrale costituita da actina filamentosa, rivestita da una guaina di altre proteine.

Un altro promettente campo di ricerche, tuttavia ancora in embrione, riguarda le differenze strutturali del reticolo microtrabecolare in cellule normali e cancerose. Il reticolo delle seconde è più denso di quello delle prime e mostra segni di una mancanza di organizzazione normale, particolarmente nella distribuzione dei polisomi.

In conclusione, la sostanza fondamentale del citoplasma, per molto tempo ritenuta una soluzione omogenea, ricca di proteine, viene oggi considerata come una struttura estremamente intricata, con un proprio comportamento altrettanto complesso. Il reticolo microtrabecolare organizza i diversi componenti della cellula in una unità funzionale - il citoplasto - e media il trasporto regolato e diretto all'interno della cellula. Sta anche diventando sempre più evidente che il controllo della forma e della motilità cellulari dipende dal funzionamento integrato dei microtubuli, dei microfilamenti, dei centri di organizzazione dei microtubuli e del reticolo microtrabecolare. È stata così chiaramente identificata una nuova frontiera nello studio della struttura e della fisiologia della cellula.

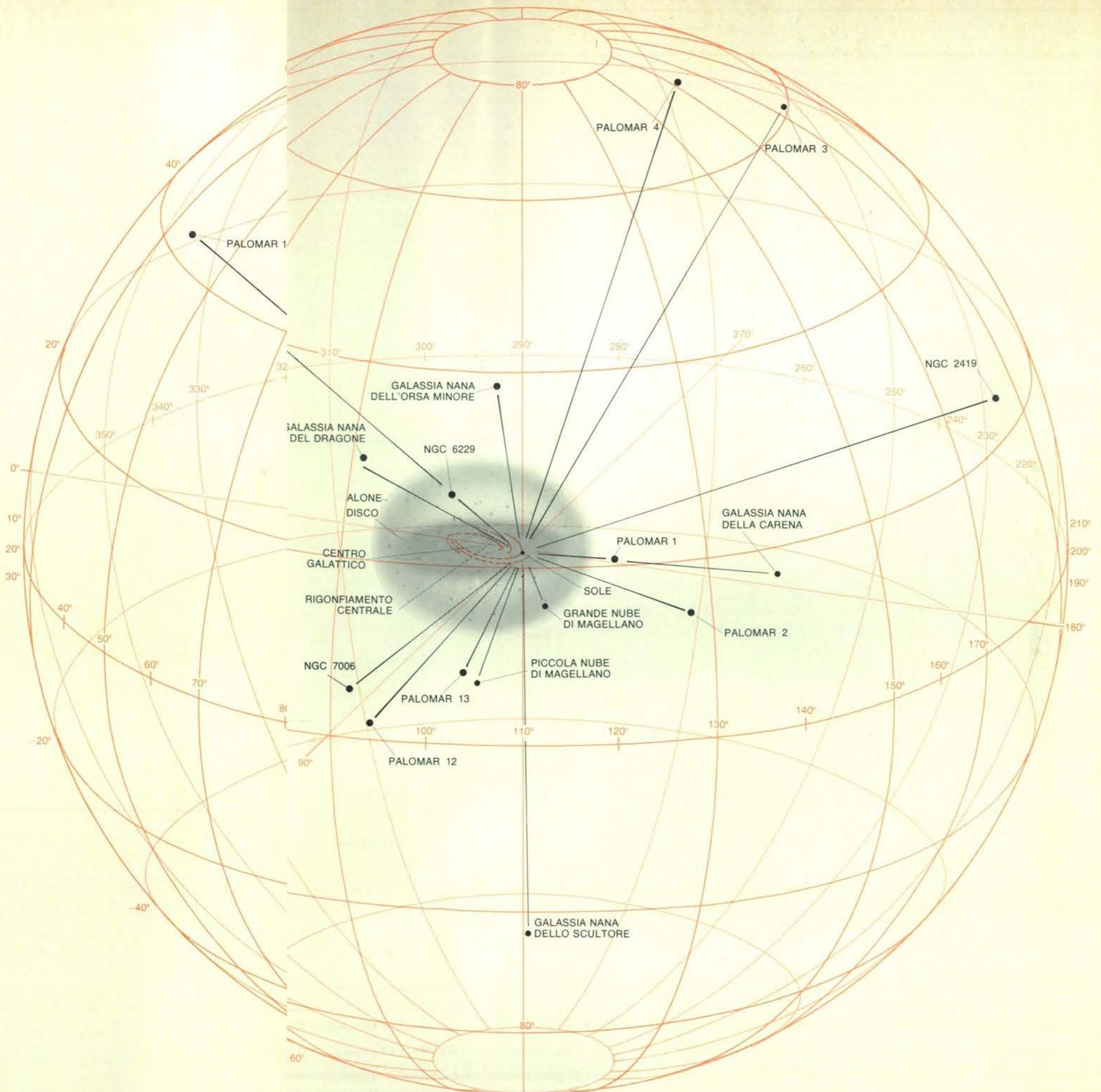
La Via Lattea

Pochi anni fa le sue principali caratteristiche sembravano ben definite, mentre oggi si pongono in discussione persino la sua massa e il suo raggio

di Bart J. Bok

In una notte chiara e senza luna, lontano dalle luci della civiltà, ciò che colpisce di più, nel cielo, è la fascia luminosa della Via Lattea. Anche senza telescopio si possono dire molte cose su essa. Per esempio, si può notare che le stelle diventano più numerose quando dirigiamo lo sguardo attraverso il cielo all'interno della fascia. Il fatto che la stessa fascia sia costituita principalmente di stelle sembra poi meno sorprendente. Si può notare che la fascia segue un grande cerchio che divide in due la sfera celeste. La Terra è quindi immersa nel piano centrale della fascia. Si può inoltre notare che questa ha la massima ampiezza e la massima luminosità in direzione della costellazione del Sagittario, che sicuramente è la direzione che punta al centro del sistema. Il sistema è la Via Lattea.

Un telescopio rivelerà che certi punti dall'apparenza di stella della Via Lattea sono in realtà grandi agglomerati di stelle. Si tratta di oggetti chiaramente lontani: sono chiamati ammassi globulari. Tra il 1918 e il 1921 Harlow Shapley impiegò i telescopi del Mount Wilson Observatory per dimostrare che gli ammassi, come le stelle della fascia, sono molto più frequenti nel Sagittario. In una zona del Sagittario che costituisce soltanto il 2 per cento del cielo, Shapley rappresentò un terzo di tutti gli ammassi globulari allora conosciuti. Evidentemente, quindi, il sistema solare è lontano dal centro della Galassia; la distanza dal Sole al centro è oggi valutata in 8500 parsec. (Un parsec è pari a 3,26 anni luce.) Da allora si è fatta strada l'idea che il sistema solare, insieme alla restante massa del piano centrale della Galassia, ruoti attorno al centro. Il Sole ruota a una velocità di circa 230 chilometri al secondo e compie un giro completo ogni 200 milioni di anni.

Nel 1930 Robert J. Trumpler del Lick Observatory dimostrò che un mezzo interstellare composto di gas e di polvere oscura la luce delle stelle, in particolare di quelle del piano centrale della Galassia. Nel 1951 William W. Morgan del Yerkes Observatory e i suoi allievi Donald E. Osterbrock e Stewart L. Sharpless fornirono le prove che la fascia è una veduta laterale dalla Terra della struttura a spirale della Galassia. Altre prove della forma a spirale vennero ben presto alle lunghezze d'onda radio. Negli anni sessanta Chia Chiao Lin del Massachusetts Institute of Technology e Frank H. Shu, allora all'Harvard College Observatory, suggerirono che nel mezzo interstellare onde di densità crescente fanno scomparire la forma a spirale e sono responsabili della formazione di stelle. Nel frattempo l'astronomia all'infrarosso fornì un metodo per esplorare l'interno di nubi interstellari oscure. Oggi pare che le nubi si trovino nei punti nei quali si forma la maggior

Una mappa della Via Lattea presenta la Galassia secondo l'ipotesi che essa sia inaspettatamente grande e massiccia. Le tre componenti della Galassia da tempo note si trovano ben all'interno del reticolo di coordinate. Le loro dimensioni sono espresse in parsec. (Un parsec è pari a 3,26 anni luce.) Delle tre componenti il rigonfiamento centrale, con un raggio compreso tra 4000 e 5000 parsec, è formato principalmente da una densa concentrazione di stelle vecchie. Il disco galattico, con un raggio di 15 000 parsec, è formato da stelle più giovani, da polvere e da gas. La sua forma a spirale è stata tratteggiata solo in prossimità del Sole *(curve in colore)*. L'alone galattico, con un raggio di 20 000 parsec, è formato principalmente da una debole concentrazione di stelle vecchie e da quasi metà degli aggregati stellari chiamati ammassi globulari. L'ipotetica componente più esterna della Galassia è chiamata corona; la sua presenza è dedotta dalle velocità della materia visibile. È probabile che gli oggetti della corona non siano altamente luminosi. In questa mappa le «compagne galattiche» visibili nella corona a una distanza di 100 000 parsec sono rappresentate graficamente con coordinate aventi l'origine nel Sole. La congiungente il Sole con il centro della Galassia definisce zero gradi di longitudine galattica. Gli angoli al di sopra e al di sotto del piano galattico sono misurati in latitudine galattica. Le compagne comprendono 10 ammassi globulari, quattro galassie sferoidali nane e le galassie irregolari Grande e Piccola Nube di Magellano.

parte delle nuove stelle della Galassia. Fin dal 1970 i radioastronomi hanno iniziato a rivelare la composizione delle nubi oscure: queste sono formate da polvere e da idrogeno, con l'aggiunta di alcune molecole sorprendentemente complesse, molte delle quali organiche.

Per molti anni sono stato un osservatore notturno della Via Lattea. Ricordo la metà degli anni settanta come un periodo in cui io e i miei colleghi di osservazione ci sentivamo particolarmente sicuri. I larghi contorni della Galassia parevano ragionevolmente ben definiti. La Galassia aveva due componenti principali: un denso rigonfiamento centrale di stelle con un confine da 4000 a 5000 parsec dal centro e un disco molto più sottile e piatto, di stelle e di gas e polvere interstellare,il cui bordo interno confinava con il rigonfiamento e il cui bordo esterno si trovava a circa 15 000 parsec dal centro. Il rigonfiamento è nel Sagittario, mentre il disco si allunga nel cielo.

Allora si stimava che la massa complessiva del disco e del rigonfiamento fosse di molto inferiore a 200 miliardi di volte la massa del Sole. Tuttavia, il disco e il rigonfiamento erano circondati da un «alone» di materia da entrambi i lati del piano centrale della Galassia. Nella sua forma complessiva l'alone è una sfera leggermente schiacciata. Esso ha nel piano della Galassia un raggio dell'ordine di 20 000 parsec, ed è notevole per la presenza di stelle vecchie e anche per la diffusione di circa 100 ammassi globulari. (Altri 100 ammassi globulari si trovano all'interno o nelle vicinanze del disco galattico.) L'alone potrebbe aggiungere al più 100 miliardi di masse solari alla massa totale della Galassia.

Si doveva lavorare ancora molto. Tuttavia confidavamo che le scoperte fondamentali fossero valide e che gli astronomi non si sarebbero distratti dall'analisi di argomenti quali la formazione delle stelle e l'evoluzione delle forme a spirale della Via Lattea. Non sospettavamo che presto sarebbe stato necessario rivedere i valori accettati per il raggio e la massa della Via Lattea aumentando il raggio di un fattore tre o più, e la massa addirittura di un fattore 10. Le revisioni sono la conseguenza di un certo numero di recenti scoperte. Esaminerò in questo articolo numerosi aspetti degli sforzi in corso diretti a una maggiore comprensione della Via Lattea. Ne includerò molti perché sono fonte di notevole fermento che potrà presto portare a un progresso delle nostre conoscenze. Parlerò poi anche di un altro aspetto che, al contrario, temo stia arrivando a un punto morto.

Corone galattiche

Qualcuno cominciò a pensare che la Via Lattea fosse più grande e massiccia del previsto, già dal 1974. Dal punto di vista teorico le nostre ipotesi vennero principalmente da Donald Lynden-Bell dell'Università di Cambridge e da Jeremiah P. Ostriker, P. J. E. Peebles e Amos Yahil della Princeton University. L'argomentazione fondamentale era che la stabilità dinamica e la permanenza della Galassia non possono essere garantite a meno che il disco galattico non sia circondato, e quindi stabilizzato gravitazionalmente nonostante la sua sottigliezza e la sua delicata struttura a spirale, da un alone esteso e pesante.

J. Einasto e i suoi colleghi dell'Osservatorio Tartu in Estonia fecero un ragionamento diverso. Per molti anni i ricercatori che studiavano la dinamica della Via Lattea avevano dedotto che la velocità del Sole rispetto agli ammassi globulari dell'alone della Galassia fosse di soli 180 chilometri al secondo circa. Dal momento che gli ammassi sono distribuiti in un grande volume sferico, sembra chiaro che essi non partecipano nel loro insieme ad alcuna rotazione che il disco galattico possa avere. La distribuzione degli ammassi dovrebbe quindi essere più o meno stazionaria rispetto al centro della Galassia. In particolare, si ritiene che la velocità di rotazione del sistema degli ammassi globulari attorno al centro della Via Lattea non possa superare i 50 chilometri al secondo. Quindi la velocità di rotazione del Sole attorno al centro della Galassia non dovrebbe superare i 230 chilometri al secondo, e sicuramente non i 250.

Nel frattempo altri ricercatori avevano determinato la velocità del Sole rispetto al moto medio delle galassie vicine, cioè quelle la cui distanza dal Sole non supera un milione di parsec. Il loro risultato, per la velocità di rotazione del Sole attorno al centro della Via Lattea, fu di 300 chilometri al secondo. La differenza tra i due risultati si potrebbe benissimo interpretare come una prova che il centro della Via Lattea si sta muovendo con una velocità compresa fra 50 e 80 chilometri al secondo rispetto alle galassie vicine. Tale velocità pareva sorprendentemente grande.

Einasto pensò che il risultato poteva essere egualmente corretto, e che la causa potesse risiedere nell'insospettata massa complessiva del sistema della Via Lattea. Per verificare tale ipotesi egli esaminò il moto della Via Lattea all'interno del gruppo di galassie del quale fa parte. Einasto avanzò l'ipotesi dell'esistenza di sottogruppi all'interno del gruppo; in uno dei sottogruppi la Via Lattea avrebbe il predominio gravitazionale su numerosi altri agglomerati di stelle, quali le due piccole galassie vicine chiamate Grande e Piccola Nube di Magellano e un certo numero di galassie sferoidali nane, delle quali se ne conoscono sette. Una delle galassie sferoidali nane si trova a circa 150 000 parsec dal centro della Via Lattea. La velocità del Sole rispetto al moto medio di queste compagne galattiche era di quasi 300 chilometri al secondo, anch'essa sorprendentemente elevata. Einasto interpretò la velocità come un effetto prodotto dalla grande massa della nostra Galassia sulle galassie più vicine.

Einasto vedeva in tal modo rinforzati i suoi dubbi sul fatto che la Via Lattea fosse più estesa e più pesante di quanto si supponesse. Nel 1976 presentò un modello del sistema della Via Lattea nel quale la massa del rigonfiamento centrale, del disco e di un esteso alone, è di 900 miliardi di masse solari. Perfino tale massa è insufficiente per spiegare le grandi velocità osservate tra la Galassia e le sue compagne. Perciò Einasto ipotizzò che il rigonfiamento, il disco e l'alone fossero immersi in una componente della Galassia ancora più grande ma ciononostante non osservata, la corona (come egli la chiamò) che si estende all'esterno per almeno 100 000 parsec dal centro e ha una massa di 1200 miliardi di masse solari. La massa totale della Via Lattea dovrebbe essere allora di 2100 miliardi di masse solari, almeno sette volte il valore accettato nel 1975.

Numerose ricerche fornirono prove a sostegno di tale ipotesi. Dapprima Vera C. Rubin, W. Kent Ford, Jr., e Norbert Thonnard del Dipartimento di magnetismo terrestre della Carnegie Institution di Washington esaminarono gli spostamenti Doppler delle righe spettrali della luce emessa dalla materia della parte esterna di 17 galassie. Ciascuno di tali spostamenti è uno spostamento di una riga spettrale a una lunghezza d'onda diversa da quella che dovrebbe avere se la sorgente di radiazione fosse immobile rispetto allo strumento che la riceve. I ricercatori conclusero che la materia esterna di ogni galassia ruota attorno al centro della galassia con la stessa velocità della materia più vicina al centro. In altre parole, la curva di rotazione - un grafico della velocità di rotazione in funzione della distanza dal centro della galassia - era in pratica una retta orizzontale per la parte esterna di ogni galassia da essi studiata.

La scoperta di tali curve di rotazione piatte è notevole. Dopo tutto, la distribuzione di luminanza nelle immagini ottiche di una tipica galassia a spirale ci induce a concludere che la materia visibile della galassia sia concentrata verso il centro e più diffusa alla periferia. Da questa concentrazione si può dedurre che la materia visibile più esterna dovrebbe essere in moto a causa di forze analoghe a quelle agenti sui pianeti più esterni del sistema solare. La materia periferica dovrebbe

La curva di rotazione rappresenta graficamente la velocità periferica della materia in rotazione attorno al centro della Via Lattea. Sono qui disegnate due di tali curve. Una curva di rotazione rappresentata graficamente nel 1965 da Marten Schmidt degli Hale Observatories *(linea continua in nero)* mostra una velocità periferica che diminuisce verso il limite della Galassia visibile a 20 000 parsec. Se tutta la massa della Galassia si trova all'interno di tale limite, una massa di prova posta al di là dovrebbe ruotare a velocità inferiore, seguendo approssimativamente una legge formulata per la prima volta (per il moto dei pianeti) da Johannes Kepler *(linea tratteggiata)*. I dati analizzati da Leo Blitz e dai suoi colleghi dell'Università della California a Berkeley forniscono oggi una curva di rotazione *(linea in colore)* che sale verso un valore di 300 chilometri al secondo a 20 000 parsec. Dalla crescita della curva più recente si deduce l'esistenza di una massa invisibile di grande entità all'esterno del limite visibile della Galassia. Ogni punto dei dati più recenti è la velocità periferica di una nube di atomi di idrogeno.

La galassia nana della Carena è l'ultima arrivata nel catalogo delle compagne galattiche; la sua distanza stimata di circa 100 000 parsec dal centro della Via Lattea la rende in apparenza un membro esterno della corona galattica. In questa stampa fotografica in negativo, la galassia nana è il piccolo e modesto agglomerato di piccoli punti neri vicini al centro. I punti più grandi sono stelle molto più vicine al sistema solare lungo la stessa linea di vista. La fotografia è stata ripresa durante lo svolgimento del programma United Kingdom-European Southern Observatory Sky Survey.

Il diagramma di Hertzsprung-Russell per stelle della parte centrale della galassia nana Carena permette di determinare la distanza di tale compagna galattica. Il diagramma rappresenta graficamente la luminosità apparente di ognuna di tali stelle *(asse verticale)* in funzione di una misura del suo colore *(asse orizzontale)*. Le stelle di quello che è chiamato ramo orizzontale del diagramma risultano avere una luminosità apparente di circa 20,4. La loro luminosità intrinseca le porterebbe a magnitudini 20 volte maggiori. Dall'offuscamento della loro luminosità si può calcolare una distanza di 100 000 parsec. A una latitudine galattica di -22 gradi, Carena è ben all'esterno del piano centrale della Via Lattea, quindi la correzione dei calcoli per l'assorbimento della luce dovuto a materia stellare intermedia dovrebbe essere piccola. I dati sono stati raccolti da Russell Cannon e collaboratori al Royal Observatory di Edimburgo e all'Anglo-Australian Observatory. Il limite di sensibilità dell'apparecchiatura usata è appena al di sotto del ramo orizzontale.

La parte centrale della Via Lattea è mostrata schematicamente in questa illustrazione ideata da Thomas R. Geballe degli Hale Observatories. La vista è dall'alto. La scala diminuisce logaritmicamente con la distanza dal centro e quindi i parsec più interni sono ingranditi. I tre parsec più centrali comprendono la più densa concentrazione di stelle *(punti in colore)* della Galassia. (In tutta la figura la densità delle stelle è stata ridotta di un fattore 2000.) La regione comprende anche nubi compatte di gas ionizzato *(in colore intenso)* costituite principalmente da idrogeno. La velocità delle nubi suggerisce che siano in rotazione attorno a un oggetto supermassiccio, forse un buco nero. Una distribuzione più rarefatta di gas ionizzato *(in colore più chiaro)* riempie i 100 parsec centrali; essa è circondata da un anello di gas più freddo non ionizzato *(in grigio)*, il cui idrogeno è composto sia da atomi sia da molecole. L'anello comprende complessi giganti di polvere e di molecole *(in grigio scuro)*. In alcuni si sono formate giovani stelle *(punti bianchi)*. A sinistra appare una densa banda formata principalmente da idrogeno non ionizzato. È la parte più interna di una forma in espansione chiamata braccio da tre chiloparsec, che si pensa sia stata prodotta da un'esplosione nel centro galattico 30 milioni di anni fa.

essere in rotazione attorno al centro della galassia a una velocità periferica inferiore (espressa in chilometri al secondo) a quella della materia vicina al centro. Evidentemente, non lo è. Le curve di rotazione calcolate da Rubin, Ford e Thonnard implicano quindi la presenza di materia inosservata in grande quantità al di là dell'apparente periferia di ogni galassia.

La corona della Via Lattea

Studi analoghi a quelli di Rubin, Ford e Thonnard riguardano direttamente la distribuzione di massa della Via Lattea. F. D. A. Hartwick dell'Università di Victoria e Wallace L. W. Sargent del California Institute of Technology calcolarono le velocità di ammassi globulari a distanze maggiori di 20 000 parsec dal centro della Galassia. James E. Gunn, Gillian R. Knapp e Scott D. Tremaine utilizzarono dati raccolti all'Owens Valley Radio Observatory per determinare le velocità di nubi di atomi di idrogeno interstellari. Maurice P. Fitzgerald dell'Università di Waterloo, insieme a Peter D. Jackson dell'Università del Maryland e ad Anthony Moffat dell'Università di Montreal, determinò le velocità di stelle e di ammassi stellari fino a 17 000 parsec dal centro della Galassia. William L. H. Shuter dell'Università della Columbia Britannica determinò le velocità di nubi di idrogeno e di monossido di carbonio che si trovano nell'anticentro galattico: la direzione opposta alla linea di vista che va dalla Terra al centro della Galassia. Recentemente Leo Blitz e i suoi colleghi dell'Università della California a Berkeley hanno misurato gli spostamenti Doppler sia nella parte ottica che in quella radio dello spettro elettromagnetico per righe di 184 nebulose e di grandi nubi stellari di idrogeno e di monossido di carbonio interstellari dell'anticentro. Tutti i risultati sono coerenti con una curva di rotazione senza diminuzione.

Dai dati oggi disponibili sembra corretto concludere che la curva di rotazione della Via Lattea raggiunge un valore di 230 chilometri al secondo a 8500 parsec, la distanza dal centro galattico che contraddistingue la posizione del Sole. Da tale posizione la velocità di rotazione continua ad aumentare, per raggiungere 300 chilometri al secondo a una distanza di 20 000 parsec.

Molti oggetti più lontani di 20 000 parsec sono visibili; un elenco di essi è stato pubblicato dall'Unione astronomica internazionale nel 1979. Quattro ammassi globulari si trovano tra 20 000 e 40 000 parsec dal centro galattico; la Grande Nube di Magellano e due ammassi globulari si trovano tra 40 000 e 60 000 parsec; due galassie nane sferoidali e la Piccola Nube di Magellano si trovano tra 60 000 e 80 000 parsec; una galassia nana sferoidale e tre ammassi globulari si trovano tra 80 000 e 100 000 parsec. Altre quattro galassie nane sferoidali e due ammassi globulari si trovano tra 100 000 e 220 000 parsec; ma la loro appartenenza alla corona della Via Lattea è più incerta.

Evidentemente la nostra galassia è molto più estesa e di massa molto maggiore di quanto si ritenesse; la Via Lattea si è elevata al rango di una grande galassia a spirale. Nel loro insieme, tuttavia, i costituenti visibili della corona sono solo una piccolissima frazione delle centinaia di miliardi di masse solari calcolate. In apparenza la Via Lattea ha in comune con le galassie studiate da Rubin, Ford e Thonnard la proprietà che la maggior parte della sua materia esterna è oscura. Infatti, John N. Bahcall e Raymond M. Soneira dell'Institute for Advanced Study concludono dall'invisibilità dell'ipotetica materia che se nella corona esistono stelle non legate in ammassi o in galassie nane, la loro luminanza intrinseca dovrebbe essere inferiore a un millesimo di quella del Sole.

Da che cosa è formata allora la massa non osservata? Tre fatti sono degni di nota. In primo luogo, le galassie nane e gli ammassi globulari sono formati principalmente da stelle vecchie. In secondo luogo, le stelle vecchie non sono altamente luminose. In terzo luogo, nessuno ha rivelato dalla corona le righe spettrali che caratterizzano nubi di materia gassosa quale l'idrogeno e il monossido di carbonio in parti più centrali della Galassia. Al momento attuale, quindi, l'ipotesi più attendibile è che la corona della Via Lattea sia composta principalmente da stelle vecchie. Del resto, la massa non osservata della corona della Galassia non può adattarsi ad alcuna delle categorie basate su ciò che si può osservare in regioni più accessibili. Non sappiamo ancora che cosa c'è al di fuori.

Il rigonfiamento centrale

Può sembrare sorprendente il fatto che il centro della Via Lattea sia non meno misterioso della corona galattica. Per osservatori sulla Terra, infatti, è solo di poco più visibile. Venticinque anni fa, quando si potevano effettuare osservazioni solo alle lunghezze d'onda visibili e radio, si conosceva l'abbondanza nel centro galattico di tre tipi di oggetti, tutti vecchi. I primi sono gli ammassi globulari, i secondi le variabili RR Lyrae (vecchie stelle che si accendono e si spengono alternativamente con un periodo dell'ordine di un giorno). I terzi sono le nebulose planetarie, che per la maggior parte sono le stelle vecchie collassate dette nane bianche, ognuna delle quali è circondata da una nube di gas che si pensa sia l'atmosfera di schermo della stella. Lo stesso centro galattico era nascosto all'osservazione ottica da strati e involucri di polvere che, secondo i calcoli, attenuano la luce dal centro fino a 30 magnitudini astronomiche, pari a un fattore 10^{12}. La polvere è particolarmente efficace nell'assorbimento della luce blu, proprio quella cioè a cui spesso erano più sensibili le emulsioni fotografiche impiegate agli inizi della fotografia astronomica. Una parte della polvere si trova su una striscia ai margini del rigonfiamento centrale, a poche migliaia di parsec dal sistema solare. Una parte di essa si trova a soli 300 parsec dal sistema solare, nella costellazione dell'Ofiuco, lungo la linea di vista dalla Terra al

centro. Tutto quello che allora si sapeva, sul centro, era che esso ospitava un'intensa radiosorgente. La radiazione elettromagnetica nella regione radio dello spettro non viene assorbita dalla polvere.

Circa 15 anni fa le prospettive cominciarono a migliorare. La rivelazione di emissioni radio a una lunghezza d'onda di 21 centimetri servì a individuare le sorgenti di tale radiazione: nubi di atomi di idrogeno neutri (non ionizzati). La scoperta di emissioni ad altre lunghezze d'onda rivelò la presenza di nubi oscure interstellari formate principalmente da molecole. Soprattutto, l'astronomia all'infrarosso aprì una finestra sul centro della Galassia a lunghezze d'onda comprese tra un micrometro e un millimetro. La radiazione infrarossa costituisce una seconda banda di lunghezze d'onda che la polvere interstellare non assorbe fortemente. Una delle nuove tecniche più utili è la rivelazione della radiazione infrarossa emessa a una lunghezza d'onda di 12,8 micrometri da atomi di neo ionizzati. Dal momento che il neo è un prodotto secondario di eventi energetici quali le esplosioni stellari, esso diventa un costituente regolare delle nubi interstellari. Il neo ha perciò la funzione di indicatore.

Le osservazioni a lunghezze d'onda radio e infrarosse forniscono quattro tipi di dati. In primo luogo, esse rivelano la presenza di massimi locali nella pioggia di radiazioni da certe regioni del cielo. In secondo luogo, lo spostamento Doppler delle righe spettrali della radiazione rivela la velocità radiale di una tale sorgente: la velocità di avvicinamento o di allontanamento dal sistema solare. In terzo luogo, l'allargamento di una riga spettrale può suggerire che la sorgente si stia espandendo o contraendo. In quarto luogo, le intensità relative di certe righe spettrali suggeriscono la temperatura della sorgente. Anche quando sono state estratte tutte le informazioni possibili dalla radiazione, tuttavia, molte cose rimangono incerte. Per esempio, non si può chiarire se una sorgente che si trova all'incirca nella direzione del centro galattico è davanti al centro o dietro di esso, o se la velocità radiale di una sorgente è realmente una prova del fatto che essa sta ruotando attorno al centro.

Riassumerò lo stato attuale delle conoscenze sul rigonfiamento centrale della Via Lattea. Nella sua forma complessiva il rigonfiamento centrale è una sfera leggermente appiattita. Il suo contorno esterno, che si trova a 5000 parsec dal centro, è caratterizzato da un anello di quelli che sono oggi chiamati complessi molecolari giganti. Essi sono grandi, buie, confuse nubi interstellari costituite principalmente da molecole di idrogeno. Ne riparlerò in seguito. Lo stesso rigonfiamento è formato in generale da un denso ammasso di stelle vecchie in una matrice piuttosto sottile di gas e polvere interstellare. Le stelle si riconoscono per la loro radiazione infrarossa, che si può distinguere da quella del gas o della polvere. Una possibile spiegazione della scarsità relativa di gas e di polvere nel rigonfiamento sta nel supporre che la maggior parte di esso si sia condensato tranquillamente molto tempo fa per formare le stelle del rigonfiamento. D'altra parte, parecchie caratteristiche interstellari del rigonfiamento fanno pensare che il centro della Galassia abbia avuto una storia complessa e violenta.

La più esterna fra le caratteristiche rivelabili del rigonfiamento è un anello di idrogeno neutro a una distanza di 3000 parsec dal centro. L'anello è stato scoperto nel 1964 da Jan H. Oort e G. W. Rougoor dell'Osservatorio di Leida. Gli spo-

La struttura a spirale del disco della Via Lattea in prossimità del Sole è stata ricavata da osservazioni a lunghezze d'onda visibili con alterna fortuna, a seconda di quali classi di oggetti astronomici si impiegano nel tentativo. La carta superiore mostra le posizioni di ristretti gruppi di giovani stelle *(cerchietti vuoti)* e di ammassi nei quali almeno alcune stelle sono giovani *(cerchietti pieni)*; la distribuzione di tali stelle conferma la presenza di forme a spirale chiamate il braccio di Perseo, il braccio di Orione e il braccio Sagittario-Carena. La carta inferiore mostra invece le posizioni delle variabili Cefeidi a lungo periodo. Si pensa che queste giovani stelle giganti formino internamente segmenti di spirale. Nonostante le due mappe disegnate corrispondano alla stessa parte della Galassia, nella seconda non appare alcuna struttura a spirale. I dati utilizzati per la compilazione delle carte sono stati raccolti da Roberta M. Humphreys dell'Università del Minnesota.

La più luminosa formazione a spirale in prossimità del Sole è la Grande Nebulosa della Carena, che individua il punto in cui si incontrano il braccio del Sagittario e quello della Carena. La nebulosa è una nube di idrogeno che è stato ionizzato e reso pertanto luminoso dalla radiazione ultravioletta delle stelle supergiganti bianco-blu di nuova formazione presenti nella sua foschia. La nebulosa, che non può essere osservata dall'emisfero settentrionale, dista 2700 parsec. La fotografia è stata ripresa al Mount Stromlo Observatory dell'Australian National University.

stamenti Doppler della radiazione che esso emette mostrano che è in rotazione e, fatto più importante, che è in espansione con velocità, che lo allontanano dal centro, comprese fra 50 e 135 chilometri al secondo. Forse l'anello è un nuovo braccio a spirale che si sta stendendo. Comunque, si è ugualmente tentati di ipotizzare che il centro della Galassia abbia espulso una specie di anello di fumo circa 30 milioni di anni fa. È come se in quel punto si fosse verificata una esplosione titanica. Forse l'esplosione ha spazzato via la maggior parte della polvere e del gas del rigonfiamento.

Procedendo verso l'interno, la caratteristica successiva si trova a una distanza di circa 1500 parsec dal centro. Gli scopritori, Butler Burton dell'Università del Minnesota e Harvey S. Liszt del National Radio Astronomy Observatory, hanno avanzato l'ipotesi che si tratti di un disco di idrogeno sia atomico sia molecolare. Anch'esso è in rotazione e in espansione. Sorprendentemente, il miglior modo di interpretare i dati è di supporre che il disco sia inclinato di un angolo, fra 15 e 20 gradi, rispetto al piano della Galassia.

Si potrebbe sperare che la composizione del rigonfiamento centrale sia più o meno omogenea a partire dall'anello di Burton-Liszt verso il centro, se non altro per facilitare il compito di quanti stanno tentando di capire la struttura del rigonfiamento. Invece, ci attendono altre sorprese. Vi sono evidenti tracce di un altro anello di fumo a circa 300 parsec dal centro. Anch'esso è un miscuglio di complessi molecolari, nubi di polvere e regioni di idrogeno atomico e molecolare. L'idrogeno atomico è ionizzato in zone ben definite, il che significa che esso è molto caldo: ben oltre i 10 000 kelvin. Insieme a queste macchie calde esistono ammassi di stelle supergiganti bianco-blu di nuova formazione. Per quale motivo queste regioni di alta temperatura e di formazione di stelle si trovano in questo anello? Si resta particolarmente incerti dal momento che un anello più freddo di atomi ionizzati solo moderatamente, a una temperatura di 5000 gradi, si trova a soli 10 parsec dal centro. L'anello a 10 parsec è piuttosto denso e sta ruotando.

I tre parsec centrali della Galassia comprendono evidentemente parecchi milioni di stelle, che costituiscono il centro del più denso accumulo di stelle della Galassia. La regione del nucleo comprende anche un certo numero di nubi compatte di gas ionizzato; un gruppo di ricercatori guidato da John H. Lacy e da Charles H. Townes dell'Università della California a Berkeley ne ha rivelate 14. Una tipica nube ha all'incirca la stessa massa del Sole, un diametro di una frazione di parsec e ruota attorno al centro, compiendo un'orbita in circa 10 000 anni (in confronto al periodo orbitale del Sole, che è di circa 200 milioni di anni). Luis Rodriguez ed Eric J. Chaisson dell'Harvard College Observatory hanno mostrato che la velocità del gas ionizzato aumenta in prossimità del centro. Tutto ciò fa pensare che le nubi siano satelliti di un oggetto più interno superpesante.

Qualunque cosa si trovi nel centro stesso, appare come una brillante sorgente infrarossa nelle mappe disegnate da Eric E. Becklin e Gerry Neugebauer del California Institute of Technology. Secondo Bruce Balick dell'Università di Washington e Robert L. Brown del National Radio Astronomy Observatory, che hanno studiato le emissioni radio dell'oggetto centrale, esso ha un diametro non superiore a 10 volte la distanza tra la Terra e il Sole. La sua massa può raggiungere i 50 milioni di masse solari. L'ipotesi più probabile è che nel centro stesso della Via Lattea esista un buco nero creato dall'accumularsi di centinaia di migliaia di stelle. Il centro dovrebbe essere in ultima analisi un cimitero stellare.

Il più ricco ammasso globulare in prossimità del Sole è l'Omega Centauri, qui riprodotto in una stampa fotografica in negativo ripresa da Gary S. Da Costa con il telescopio da 2,5 metri dell'Osservatorio Las Campanas in Cile. L'ammasso comprende parecchie centinaia di migliaia di stelle entro un diametro di circa 30 parsec. Omega Centauri si trova a poco più di 5000 parsec di distanza dal sistema solare.

La curiosa composizione stellare dell'ammasso globulare della Omega Centauri è suggerita da un diagramma di Hertzsprung-Russell per le stelle che formano l'ammasso. Il ramo del diagramma comprendente stelle rosse giganti *(in alto a destra)* è insolitamente lungo. Ciò indica un insolito grado di variazione nel contenuto di «metalli», ossia di atomi più pesanti del carbonio, delle stelle giganti. Più esattamente, gli atomi metallici della stella aumentano la sua opacità alla radiazione che sfugge dal suo interno. Esse pertanto cambiano colore: le giganti rosse ricche di metalli sono più rosse e, in un diagramma di Hertzsprung-Russell, la loro posizione è nettamente a destra. La variazione del contenuto metallico fa pensare a sua volta che le stelle dell'ammasso dell'Omega Centauri si siano formate in più riprese. Ulteriori conclusioni sono problematiche. In generale l'età di un ammasso globulare si determina confrontando la luminosità e il colore delle stelle in corrispondenza a quello che è chiamato taglio della sequenza principale del diagramma di Hertzsprung-Russell *(in basso)* con quelli delle stelle nella corrispondente posizione nei diagrammi ricavati in simulazioni con elaboratore dell'evoluzione di un ammasso. Per Omega Centauri i dati non rivelano la completa estensione della sequenza principale e risulta difficile individuare esattamente la posizione del taglio. Ciò rende difficile stabilire l'età (o le età) dell'ammasso. I dati sono stati raccolti da Cannon e da N. J. Stewart del Royal Observatory di Edimburgo.

Struttura a spirale nell'ottico

Una volta esaminati la periferia e il centro della Via Lattea, rimane una regione centrale, al cui centro si trova il sistema solare. È la parte di Galassia nella quale prevale la struttura a spirale. Le prime rappresentazioni dei bracci a spirale risalgono a trent'anni addietro, quando Morgan, Osterbrock e Sharpless individuarono tre segmenti di bracci a spirale. Per disegnarli essi rappresentarono graficamente le posizioni delle stelle supergiganti bianco-blu classificate sulla base del loro schema di righe spettrali come stelle *O* e *B*, insieme alle nubi luminose di atomi di idrogeno ionizzati che circondano spesso tali stelle. Tali diagrammi mostrano un braccio di Orione, del quale fa parte il Sole; un braccio di Perseo, 2000 parsec più lontano dal centro della Galassia, e un braccio del Sagittario, 2000 parsec più vicino al centro della Galassia. La denominazione dei bracci riflette la prassi diffusa in astronomia di utilizzare le costellazioni per caratterizzare certe direzioni nel cielo. Ai tre bracci individuati da Morgan e dei suoi colleghi si è aggiunto da allora il braccio della Carena, che può essere il proseguimento del braccio del Sagittario; l'insieme è chiamato braccio del Sagittario-Carena. I due segmenti che lo compongono si incontrano nella Grande Nebulosa della Carena, che comprende un gran numero di stelle *O* e *B*. Esistono prove dell'esistenza di altri segmenti.

Una recente ricerca compiuta da Roberta M. Humphreys dell'Università del Minnesota conferma che le stelle *O* e *B* sono abbondanti nei bracci principali individuati trent'anni fa. È una buona notizia, perché vuol dire che i bracci a spirale sono davvero delineati dalle caldissime stelle supergiganti bianco-blu, dagli ammassi formati da tali stelle e dalle nubi luminose di gas nelle quali tali stelle e ammassi si trovano. Le stelle *O* e *B* sono molto giovani; quelle che si osservano oggi nei segmenti di bracci a spirale della Via Lattea si sono formate non più di 10 milioni di anni fa. È fuor di dubbio che i bracci a spirale sono regioni di formazione di stelle.

Pare che i bordi interni della maggior parte dei bracci a spirale siano caratterizzati da nebulose oscure: fredde nubi di atomi, molecole e polvere. Secondo la teoria dell'onda di densità di Lin e Shu, le nebulose oscure indicano la compressione della materia interstellare da parte di un'onda di pressione che avanza attraverso il disco galattico. La compressione accelera la condensazione della materia nelle stelle del segmento della spirale. Sembrerebbe probabile che le stelle giovani si trovino dietro l'onda di densità che sta avanzando e le nebulose oscure davanti. Secondo la teoria, però, l'onda viaggia solo con i due terzi della velocità di rotazione del disco galattico. Quindi l'onda viene raggiunta dalle stelle formatesi nel suo interno e le nebulose oscure finiscono col trovarsi lungo il bordo di uscita delle stelle.

Alcuni dei risultati di Humphreys sono però meno incoraggianti. Alcune delle stelle

chiamate variabili Cefeidi sono giovani come le stelle *O* e *B*; in particolare le variabili Cefeidi che si spengono e si accendono alternativamente con un periodo maggiore di 15 giorni hanno un'età non superiore ai 10 milioni di anni. Inoltre, esse sono grandi e luminose e si possono perciò vedere a grandi distanze. Dal momento che sono giovani, non dovrebbero essersi allontanate troppo dal luogo di nascita, probabilmente una nebulosa oscura sul bordo interno di un braccio a spirale. Tutto sommato, le variabili Cefeidi a lungo periodo dovrebbero essere eccellenti indicatori della struttura a spirale, mentre invece non lo sono. Le variabili Cefeidi rappresentate graficamente da Humphreys presentano una distribuzione di punti essenzialmente casuale nel piano galattico.

Va sempre tenuto presente da parte di coloro che studiano la struttura a spirale della Via Lattea nell'ottico che il metodo usuale di calcolare la distanza di una stella è quello di confrontare la sua luminosità osservata con quella che si suppone sia la sua luminosità intrinseca. La luminosità intrinseca è determinata sulla base dello schema delle righe nello spettro della radiazione della stella. La diminuzione della luminosità è correlata alla distanza. Tuttavia, anche con le migliori tecniche, la distanza di una stella *O* o di una stella *B* è nota con un'approssimazione del 10 per cento. Per esempio, una stella che si è calcolato sia a 2700 parsec dal Sole - per esempio una stella della nebulosa della Carena - potrebbe distare sia 2400 sia 3000 parsec. Il risultato provoca un'indeterminazione puramente osservazionale di parecchie centinaia di parsec nella rappresentazione grafica di quella che potrebbe essere in realtà una forma a spirale.

Con i più moderni telescopi, spettrografi e attrezzature fotografiche si possono rivelare stelle *O* e *B* e se ne possono esaminare gli spettri fino a distanze calcolate dell'ordine di 8000 parsec dal Sole. A tale distanza, però, l'incertezza nel calcolo è di più o meno 800 parsec. In tal caso una forma a spirale potrebbe benissimo non essere riconosciuta. Pare che esistano modeste prospettive di tracciare la struttura a spirale della Via Lattea con mezzi ottici al di là di 8000 parsec.

Struttura a spirale nel radio

Come vanno i tentativi di tracciare la struttura a spirale alle lunghezze d'onda radio? Anche in questo caso le prospettive sono ben poco incoraggianti. Fin dall'inizio degli anni cinquanta, quando Morgan, Osterbrock e Sharpless presentarono gli aspetti ottici fondamentali della struttura a spirale, Harold I. Ewen ed Edward M. Purcell della Harvard University avevano rivelato la radiazione da nubi interstellari di idrogeno atomico neutro a una lunghezza d'onda di 21 centimetri. In pochi anni furono disponibili le prime mappe radio della Galassia dalla quale si poteva dedurre che almeno nelle parti esterne del disco galattico le nubi di idrogeno atomico neutro sono disposte in forme a spirale quasi circolari.

Val la pena di analizzare alcuni particolari del modo in cui vengono raccolti e interpretati i dati. Quando un radiotelescopio rivela radiazione a una lunghezza d'onda di 21 centimetri o nelle vicinanze da una piccola regione del cielo, in realtà esso riceve radiazione da un certo numero di nubi di idrogeno atomico neutro lungo una sola linea di vista. Ogni nube ha la propria velocità di avvicinamento o di recessione rispetto al telescopio e quindi la sua radiazione ha uno spostamento Doppler caratteristico. Il risultato è che il telescopio riceve da una sola direzione un profilo di picchi e valli in un grafico che riporta l'intensità in funzione della lunghezza d'onda. La famosa riga a 21 centimetri è in realtà un insieme di righe vicinissime di varie intensità, vari gradi di allargamento e vari spostamenti Doppler. Dato che si suppone che l'idrogeno atomico neutro abbia la sua massima densità lungo i bordi della forma a spirale, pare ragionevole supporre che un picco di elevata intensità del profilo si trovi alla lunghezza d'onda corrispondente alla velocità di avvicinamento o di recessione dell'idrogeno dove la linea di vista attraversa una forma a spirale. Con l'aiuto di una curva di rotazione della Via Lattea si dovrebbe essere in grado di determinare la distanza dalla forma.

Questo modo di ragionare è semplice e piacevole, ma non sta in piedi a un attento esame. La prima complicazione è che le nubi di idrogeno atomico neutro hanno moti propri, ben separati dal moto delle forme a spirale. Tali moti indipendenti possono modificare facilmente la velocità totale di una nube fino a sei chilometri al secondo. Un'ulteriore complicazione è costituita dal fatto che la materia del disco galattico mostra correnti su grande scala. Burton e i suoi collaboratori del National Radio Astronomy Observatory hanno dimostrato che anche lievi moti di questo genere possono originare picchi di intensità nella radiazione a 21 centimetri per direzioni nelle quali la linea di vista non incrocia una forma a spirale. Inoltre, vi sono direzioni del cielo lungo le quali la velocità di avvicinamento o di recessione di una nube di gas può cambiare solo lentamente con la distanza. Il profilo a 21 centimetri può quindi presentare un picco causato da contributi dovuti alla larghezza di una sola grande massa espansa di gas di densità uniforme.

Per studiare la struttura a spirale si possono seguire numerosi altri metodi. Uno è quello di osservare le righe spettrali emesse da molecole di monossido di carbonio a lunghezze d'onda radio prossime a 2,6 millimetri. Le righe del monossido di carbonio ci conducono a una classe di nubi più fredde di quelle composte da idrogeno atomico neutro. Le nubi più fredde sono composte principalmente da idrogeno molecolare, con un modesto contenuto di monossido di carbonio e di altre sostanze. (Il principio generale è che un aumento di temperatura porta dapprima alla dissociazione delle molecole in atomi neutri e successivamente alla ionizzazione degli atomi.) William Herbst della Wesleyan University ha dimostrato che la presenza di monossido di carbonio in una nube interstellare è correlata alla presenza di polvere. Le righe del monossido di carbonio aiutano quindi a rivelare nubi di polvere e complessi di polvere distanti che potrebbero sfuggire all'osservazione.

Un'esplorazione a bassa risoluzione delle nubi di monossido di carbonio eseguita da Richard S. Cohen e Patrick Thaddeus del Goddard Institute for Space Studies e da Thomas M. Dame della Columbia University evidenzia l'esistenza di forme a spirale. In realtà, in qualche caso pare che le nubi di monossido di carbonio riempiano gli spazi tra le nubi di idrogeno atomico in alcune delle forme a spirale individuate. Ciononostante, dubito che si sarebbe scoperta la struttura a spirale della Via Lattea se i soli dati disponibili fossero stati le osservazioni radio a una lunghezza d'onda di 2,6 millimetri.

Non mi piace affatto dover esser pessimista sulle prospettive di disegnare mappe della struttura a spirale della Via Lattea al di là di 8000 parsec dal Sole, anche perché la struttura pare ben definita fino a tale distanza, ma non vedo alcuna possibilità di miglioramento nei prossimi dieci o vent'anni. Ciò non vuol dire che non esistano attualmente possibilità di studio della struttura a spirale. Tra i telescopi ottici i grandi riflettori già montati e lo Space Telescope, uno strumento che entrerà in orbita attorno alla Terra, promettono di rivelare i particolari della struttura a spirale e i moti di forme in piccola scala. Tra i radiotelescopi il Westerbork Array olandese e il Very Large Array del New Mexico, che sintetizza le immagini raccolte da un gruppo di rivelatori, consentiranno di migliorare le ricerche sulle fredde e oscure nubi delle forme a spirale e senza dubbio forniranno molte indicazioni sulle cause della struttura a spirale. Questi promettenti sviluppi si applicano però principalmente alle galassie a spirale diverse dalla nostra. Ne sono esempi molto vicini le galassie a spirale Messier 31, 33, 51, 81 e 101.

L'età degli ammassi globulari

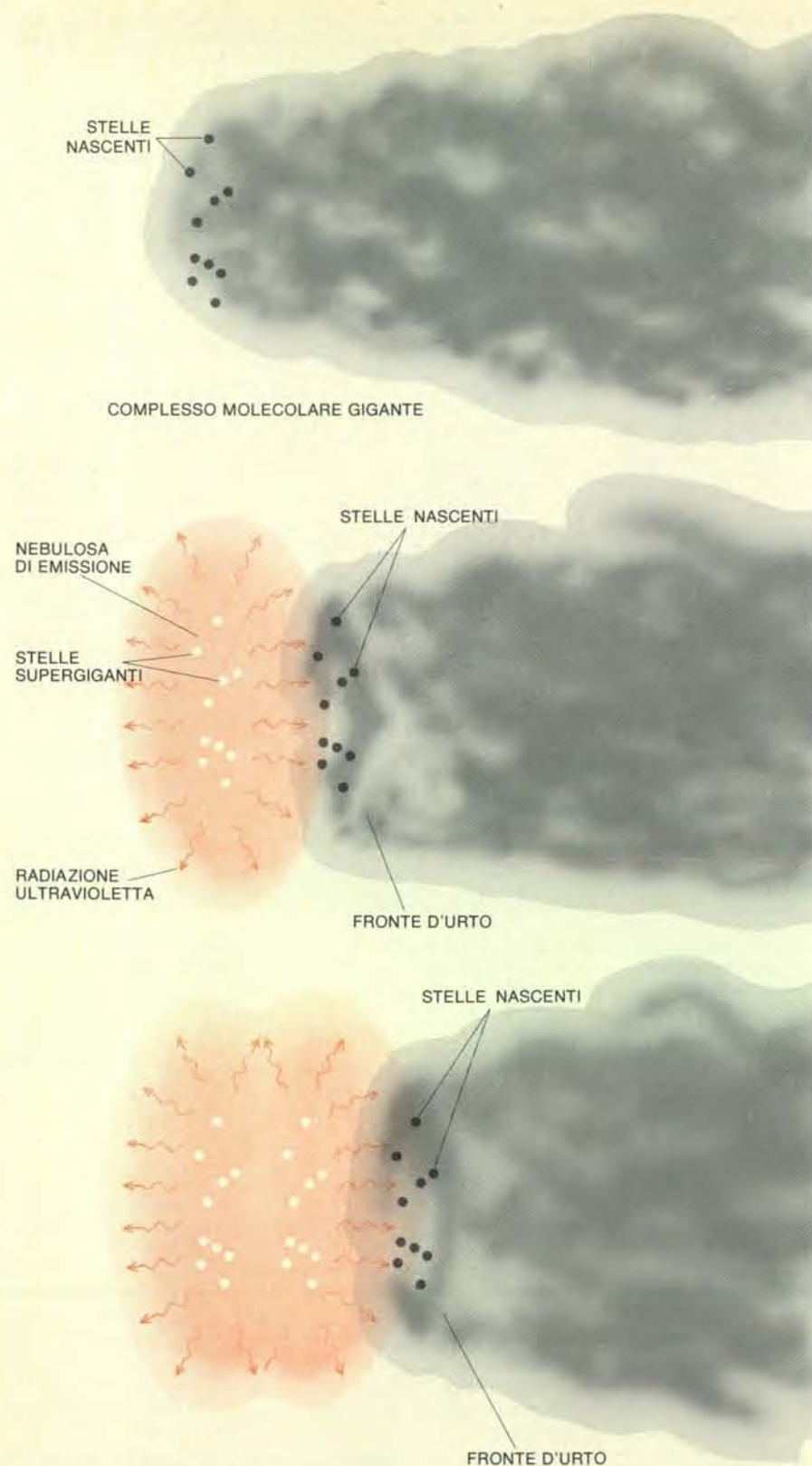
Un settore di ricerca in espansione sia in astronomia galattica sia in quella extragalattica è l'esame degli ammassi globulari. Gli ammassi sono importanti anche perché pare siano gli oggetti più vecchi della Via Lattea. Essi danno perciò informazioni sulla nascita e sull'evoluzione della Galassia e perfino del primo universo. L'ipotesi più semplice sugli ammassi è che essi si siano formati tutti in breve tempo (per esempio un miliardo di anni) durante il *big bang*: l'istante nel quale tutta la materia dell'universo attuale è comparsa in modo esplosivo da un singolo punto. Gli ammassi dovrebbero essere quindi stati tra i primi oggetti a condensare mentre le galassie prendevano forma, e ogni ammasso derivava da una grande bolla di gas. Parla a favore di tale ipotesi il fatto che metà dei circa 200 ammassi globulari della Via Lattea siano distribuiti nel volume quasi sferico dell'alone galattico. Presumibilmente essi si sono formati in tale regione ben prima che

Una curiosa nascita di stelle è visibile nella piccola nube oscura sulla sinistra di questa fotografia. La nube è una di quelle chiamate globulo. Ha un diametro di mezzo parsec e ha probabilmente una massa di poco inferiore a 100 masse solari. La massa del globulo è composta da molecole (principalmente di idrogeno) e polvere. Da questo particolare globulo è stata evidentemente espulsa una coppia di stelle nascenti collegate da una striscia luminosa. La striscia attraversa l'estremità superiore della nube. L'espulsione contrasta con l'ipotesi che un globulo condensi in circa un milione di anni per formare una sola nuova stella. Lo sfondo luminoso che si estende diagonalmente dal basso verso la parte destra del campo è il bordo della nebulosa di Gum, che si pensa sia in parte un resto di un'esplosione di supernova. Sia il globulo sia la nebulosa si trovano a circa 300 parsec dal sistema solare. Essi giacciono nella regione del cielo meridionale comprendente la costellazione della Vela. La fotografia è stata ripresa dall'autore nel 1978 con il telescopio da quattro metri del Cerro Tololo Inter-American Observatory.

prendesse forma il disco galattico. L'orbita di un tale ammasso è un'ellisse piuttosto eccentrica e non l'orbita quasi circolare caratteristica della materia del disco. L'asse maggiore del disco è talvolta di parecchie decine di migliaia di parsec e ogni miliardo di anni l'orbita spinge l'ammasso attraverso lo spessore del disco, che è di solo qualche centinaio di parsec.

Nell'ipotesi più semplice le stelle degli ammassi globulari dovrebbero essersi formate in un'epoca nella quale la materia presente nella Galassia era principalmente idrogeno ed elio, i due elementi chimici che si suppone siano stati creati immediatamente dopo il big bang. Una stella nata successivamente condenserebbe, invece, gas interstellare, parte del quale proverrebbe dal ciclo interno di stelle nate in precedenza. Per esempio, una parte del gas potrebbe essere stata scagliata nello spazio da esplosioni di supernova. Si tratterebbe di materia nella quale gli atomi più pesanti proverrebbero da fusione termonucleare. La stella nata successivamente dovrebbe quindi avere una concentrazione più elevata di elementi chimici pesanti; nel linguaggio degli astrofisici tali elementi sono tutti conosciuti come metalli.

È la concentrazione di metalli nei vari ammassi globulari che mette in pericolo la più semplice delle ipotesi. È accertato che il rapporto tra metalli e idrogeno ed elio è 100 volte maggiore nel Sole di quanto lo sia nelle stelle di ammassi globulari poveri di metalli quale M3 (per Messier 3). Inoltre, gli ammassi poveri di metalli tendono a essere quelli più esterni della Galassia. Oggi viene loro attribuita un'età di 15 miliardi di anni. D'altra parte, l'ammasso globulare 47 del Tucano è relativamente ricco di metalli. Si ritiene che la sua età sia di 10 miliardi di anni, pari al doppio dell'età del Sole. Omega Centauri, il più imponente ammasso globulare della Via Lattea, presenta una gamma di concentrazioni di metalli. Evidentemente si tratta di un caratteristico agglomerato di stelle nate in tempi differenti. L'ampia distribuzione delle età degli ammassi globulari contrasta con gli attuali modelli evolutivi della Galassia. Nessun modello ammette che la condensazione del disco galattico abbia richiesto diversi miliardi di anni.

Per quanto concerne il problema dell'età, si deve notare che la massima età dell'universo si può ricavare dalle velocità alle quali le galassie si stanno reciprocamente allontanando. Il massimo valore attuale è vicino a 15 miliardi di anni. Recenti ricerche sulla velocità di recessione tendono a ridurre tale massimo. Se continuerà l'attuale andamento, l'età potrebbe probabilmente ridursi fino a 10 miliardi di anni. In tal caso sarebbe necessario spiegare perché certi ammassi globulari abbiano apparentemente un'età superiore a quella attribuita all'universo. È molto importante che si ottengano per le età degli ammassi globulari valori attendibili e che essi siano disponibili per la fine di

Si pensa che la nascita di stelle supergiganti avvenga in fasi ripetitive. La sequenza ha inizio *(disegno in alto)* con la formazione di un ammasso di stelle supergiganti alla periferia di un complesso molecolare gigante, una regione fredda e scura di molecole (principalmente di idrogeno) e di polvere. Dopo alcuni milioni di anni la radiazione ultravioletta delle nuove stelle ha ionizzato l'idrogeno circostante e le stelle sono confuse in una nube di gas luminoso. Inoltre la pressione della radiazione ha compresso la materia in un complesso molecolare gigante. (Ogni quanto di radiazione possiede una quantità di moto e quindi spinge le molecole e la polvere.) Si ha la formazione di un secondo gruppo di stelle supergiganti *(disegno al centro)*. In alcuni successivi milioni di anni una sequenza del genere può dare origine a un ulteriore gruppo di stelle *(disegno in basso)*. Il diagramma si basa su un'ipotesi di Bruce G. Elmegreen della Columbia University e Charles J. Lada dell'Università dell'Arizona. L'ipotesi spiega gli ammassi di stelle supergiganti in nebulose, quale la Grande Nebulosa di Orione, che si trovano vicino a un complesso molecolare.

questo decennio, perché allora lo Space Telescope dovrebbe aver fornito valori di gran lunga più attendibili per le velocità di recessione e le distanze delle galassie. In tal modo si potrebbe disporre di un valore di gran lunga più attendibile sull'età massima dell'universo.

L'evoluzione degli ammassi globulari

La dinamica interna degli ammassi globulari è ben descritta da tre tempi caratteristici. Il primo è il tempo di attraversamento: il tempo impiegato da una stella ad attraversare l'ammasso sotto l'attrazione gravitazionale di tutto l'ammasso. Il secondo è il tempo di rilassamento: il tempo impiegato da una stella a sistemarsi per diventare un membro stabile dell'ammasso sotto l'azione delle sue interazioni gravitazionali con le stelle vicine. Il terzo è il tempo di evoluzione: il tempo nel quale un ammasso stabile cambia forma e composizione stellare in modo significativo. Per un ammasso globulare ricco di stelle il tempo di attraversamento è molto più breve del tempo di rilassamento, che risulta a sua volta molto più breve del tempo di evoluzione.

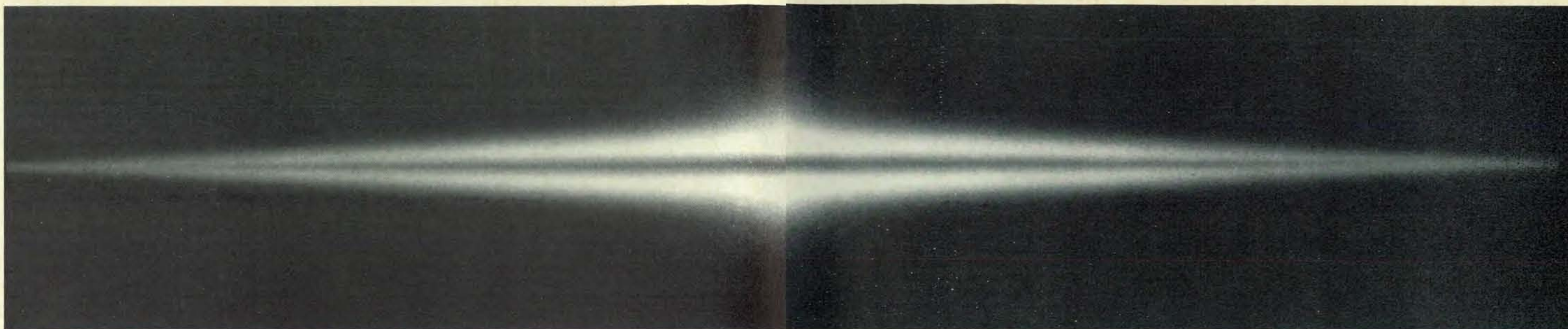
La ricerca di Ivan R. King dell'Università della California a Berkeley ha suggerito ulteriori quantificazioni. Quando un ammasso globulare ricco di stelle ha vissuto per circa 50 tempi di attraversamento, le sue stelle si sono stabilizzate; la statistica delle loro velocità sarà molto simile a quella delle velocità delle molecole in una nube di gas. Tale stato di equilibrio potrebbe persistere indefinitamente se le stelle non cambiassero mai internamente e se dall'ammasso non sfuggisse alcuna stella.

Tuttavia, fin dall'inizio degli anni trenta, Lyman Spitzer, Jr., di Princeton e Victor A. Ambartsumian dell'Osservatorio astrofisico Byurakan in Armenia avevano concluso che era molto più probabile che le stelle con massa minore di un ammasso avessero la massima velocità e che in molti casi le stelle con la massima velocità sfuggono dall'ammasso. Per la loro grande velocità le stelle che sfuggono si portano dietro più della quota che loro compete dell'energia totale dell'ammasso. Il risultato è che l'ammasso si contrae. In un periodo di tempo dell'ordine del tempo di evoluzione l'ammasso verrebbe a perdere una frazione apprezzabile dell'energia media per stella e le stelle più massicce si sistemerebbero vicine al centro.

Eventuali variazioni interne tra le stelle residue possono solo rallentare questo processo. Tanto per cominciare, la perdita di massa delle stelle è oggi ritenuta un normale fenomeno astrofisico. All'inizio della sua evoluzione, una stella tipica è immersa in un involucro gassoso che viene gradualmente espulso dalla pressione della radiazione elettromagnetica della stella diretta verso l'esterno. (Ogni fotone, o quanto della radiazione, possiede una quantità di moto e quindi spinge qualsiasi oggetto lo assorba.) Successivamente la stella comincia a emettere un leggero vento stellare di particelle. In seguito certi tipi di stelle esplodono. In una nova la stella che esplode erutta nello spazio l'equivalente della massa di un pianeta come Giove. Una supernova è una esplosione nella quale viene distrutta l'intera stella. Dal momento in cui la massa viene persa con meccanismi del genere, le stelle di un ammasso globulare si spostano sistematicamente a classi di massa inferiore (e anche di più debole luminosità intrinseca). In un periodo di tempo dell'ordine del tempo di rilassamento esse acquistano la velocità più elevata di una stella meno massiccia. Aumenta quindi la loro probabilità di fuga dall'ammasso.

Durante le centinaia di milioni di anni che un ammasso può trascorrere nel sottile alone esterno (o anche nella corona) della Via Lattea, la sua evoluzione risente soltanto degli eventi interni appena descritti. Viene però il momento in cui l'ammasso attraversa il disco galattico. In tale momento l'ammasso corre il rischio non tanto di una collisione quanto di un'interazione gravitazionale con la materia all'interno e nelle vicinanze del piano galattico. La forza di interazione può spaccare l'ammasso.

Quando l'orbita di un ammasso lo porta vicino al centro della Galassia si presenta un'altra minaccia, in quanto la massa centrale della Galassia esercita una forza attrattiva maggiore sul lato dell'ammasso che passa più vicino al centro che non sul lato che passa più lontano. La differenza tra la forza interna e quella

La nebulosa luminosa e quella oscura sono vicine nella costellazione dell'Unicorno. Esse formano insieme un esempio del modo di formazione delle stelle illustrato nella pagina a fronte. La nebulosa luminosa (o di emissione), chiamata Rosetta, è la nube oscura in alto a destra di questa negativa. I suoi atomi di idrogeno sono stati ionizzati dalla radiazione ultravioletta di un gruppo di giovani stelle supergiganti bianco-blu. Molte di tali stelle compaiono in ciò che sembra un buco al centro della nebulosa. La radiazione proveniente esattamente dal centro della stella è particolarmente intensa e può aver fatto scomparire il buco. La nebulosa oscura è un complesso molecolare gigante, formato quasi interamente da idrogeno molecolare freddo. La sua completa estensione è però meglio evidenziata dalla radiazione delle molecole di monossido di carbonio a una lunghezza d'onda radio di 2,6 millimetri. Le linee di contorno della figura evidenziano la temperatura di antenna, che indica l'intensità della radiazione del monossido di carbonio. La forma dei contorni suggerisce che la radiazione ultravioletta delle stelle supergiganti tragga nutrimento dal complesso. Pare che un fronte d'urto lungo il quale la materia del complesso viene compressa dalla radiazione si trovi sul bordo inferiore sinistro della Rosetta. Il fronte è probabilmente un luogo nel quale condenseranno nuove stelle. Un picco della radiazione del monossido di carbonio corrisponde a una sorgente di radiazione infrarossa *(IR)* oggi identificata come una stella appena nata immersa nel gas e nella polvere. Sia la nebulosa della Rosetta sia il complesso molecolare gigante si trovano a 1600 parsec di distanza dal sistema solare. La figura evidenzia i risultati di uno studio dell'interazione tra le nebulose eseguito da Leo Blitz dell'Università della California a Berkeley e da Patrick Thaddeus del Goddard Institute for Space Studies. La fotografia proviene dal Palomar Sky Survey.

La Via Lattea vista dall'esterno mostra il bordo della Galassia come potrebbe apparire al telescopio a un astronomo di un'altra galassia. L'illustrazione è stata prodotta all'elaboratore da John N. Bahcall e Raymond M. Soneira dell'Institute for Advanced Study. La simulazione è iniziata con una descrizione matematica del numero di stelle di una data brillanza in una data direzione osservate dal sistema solare. L'elaboratore ha visualizzato la figura di luminosità risultante (e una striscia centrale di polvere) come potrebbe apparire a un osservatore esterno.

esterna può deformare l'ammasso. In pratica, per qualsiasi ammasso globulare esiste un raggio critico: una distanza dal centro galattico all'interno della quale alcune stelle dell'ammasso vengono spazzate via mentre l'ammasso raggiunge la minima distanza dal centro. La distanza critica è detta «raggio di marea», perché il meccanismo che deforma l'ammasso è simile a quello che fa sollevare le maree negli oceani. Non sorprende il fatto che gli ammassi globulari all'interno del rigonfiamento centrale della Galassia siano più piccoli sia di quelli dell'alone galattico sia di quelli della corona galattica.

Un ultimo fattore può agire sugli ammassi globulari, ma lo citerò con qualche riserva. Fino a 15 anni fa pochi astronomi avrebbero detto che la formazione di sistemi di stelle doppie sarebbe stata importante negli ammassi stellari. Poi Sebastian von Hoerner del National Radio Astronomy Observatory eseguì le prime simulazioni su elaboratore di ammassi aperti: aggregati di stelle che differiscono dagli ammassi globulari nel senso che si tratta di stelle più giovani, in minor numero e più ampiamente disperse. Von Hoerner basò le simulazioni su equazioni che rappresentano le interazioni gravitazionali tra piccole quantità di stelle. Le simulazioni mostrarono chiaramente che le stelle di un ammasso tendono a unirsi in coppia per formare sistemi binari.

Lo stesso effetto è stato osservato in più recenti simulazioni all'elaboratore comprendenti dapprima 100, poi 500 e infine 1000 stelle interagenti. Nel 1975 Douglas C. Heggie dell'Università di Edimburgo dimostrò che un sistema binario di nuova formazione è un importante serbatoio dell'energia disponibile nell'ammasso. In pratica la formazione di ogni sistema binario porta via energia dal moto delle stelle per trasformarla in energia che lega le due stelle in orbita, una attorno all'altra. Di conseguenza la coppia si sposta verso il centro dell'ammasso. Ancor più di recente Spitzer e i suoi collaboratori di Princeton hanno ideato dei modelli secondo i quali nella regione centrale affollata di un ammasso globulare, nella quale si ammucchiano le stelle più massicce, è probabile che ogni stella catturi una o due o più compagne. Il risultato è che la regione centrale può diventare ancor più densamente compatta durante l'evoluzione dell'ammasso globulare.

La ragione dei miei timori nell'affrontare tutte queste prove circostanziate è semplice. In realtà i sistemi di stelle binarie predominano negli ammassi aperti. Ne è un primo esempio Mizar, una stella dell'ammasso aperto dell'Orsa Maggiore. Per meglio individuarla, dirò che si tratta della stella visibile dove il timone del Gran Carro si piega. Mizar ha una compagna chiamata Alcor, che è a malapena visibile a occhio nudo. È allora chiaro che Mizar costituisce un sistema binario. In effetti, sia Mizar sia Alcor hanno stelle compagne, e ciascuna di esse è un sistema binario. Ciò rende più esasperante il fatto che la ricerca di casi di sistemi di stelle binarie massicce vicine al centro degli ammassi globulari non sia stata ugualmente fruttifera. In realtà, essa non ha avuto alcun successo.

Si è scoperto però qualcos'altro. Nel 1976 Jonathan E. Grindlay e Herbert Gursky del Center for Astrophysics dello Harvard College Observatory e dello Smithsonian Astrophysical Observatory comunicarono che un impulso di raggi X di enorme intensità aveva raggiunto il sistema solare da una direzione prossima al centro dell'ammasso globulare NGC 6624. L'impulso ebbe una durata compresa tra otto e 10 secondi. Impulsi del genere sono stati da allora rivelati ripetutamente da sei ammassi globulari. Secondo una prima ipotesi, esiste un buco nero con una massa non inferiore a 100 masse solari nel centro di ciascuno di tali ammassi. Gli impulsi dovrebbero allora essere il risultato di collisioni intermittenti del gas interstellare con un caldo disco in espansione di gas che circonda il buco nero. Si pensa che l'ipotetico buco nero potrebbe essere l'ultima conseguenza del graduale collasso di stelle verso il centro dell'ammasso, un collasso al quale potrebbe aver contribuito la formazione di sistemi di stelle binarie, dando a esso forse anche la spinta finale.

La nascita di stelle in complessi molecolari

Era chiaro già 35 anni fa che il mezzo interstellare nel disco della Via Lattea comprende nubi di gas e polvere nelle quali stanno condensando stelle di nuova formazione. Dopo tutto, alcuni degli oggetti più interessanti della Galassia sono nebulose di emissione: luminose nubi di atomi di idrogeno ionizzati che traggono energia dai gruppi di stelle giganti al loro interno. Le stelle emettono radiazioni a un tasso talmente eccezionale da far ritenere impossibile che abbiano potuto farlo per più di alcune decine di milioni di anni. Le stelle sono quindi molto giovani. Con le sole osservazioni ottiche delle nubi era però difficile far progredire le nostre conoscenze sul vero meccanismo di formazione delle stelle.

Tutto ciò oggi è cambiato. La radioastronomia ha rivelato nelle nubi più di 50 specie di molecole interstellari, dalle molecole di idrogeno, le più leggere e le più comuni di un fattore 1000, a una catena con nove atomi di carbonio, la più pesante. Ogni specie di molecole emette radiazioni elettromagnetiche a lunghezze d'onda radio caratteristiche. Inoltre l'astronomia all'infrarosso è oggi attrezzata per rivelare perfino le stelle nascenti all'interno delle dense nubi oscuranti.

In quasi tutti i processi nei quali si ritiene si stia formando una nuova stella, la fase iniziale è la formazione di una concentrazione di materia - che chiamerò nodulo - all'interno di una nube di atomi e molecole interstellari (per la maggior parte molecole di idrogeno), con una piccola aggiunta di polvere. Alcune nubi sono in realtà enormi distribuzioni di mucchi di materia. Per esempio, un complesso molecolare gigante può avere una massa pari a centinaia di migliaia di masse solari. Altre nubi sono molto più piccole. Alcune delle nubi chiamate globuli hanno masse di sole 20 masse solari. Ciò che tutti i noduli hanno in comune è la loro tendenza a collassare, principalmente sotto l'effetto della loro stessa gravitazione, forse con pressioni occasionali dall'esterno. Una pressione esterna che può essere molto importante è quella esercitata su una nube e sui noduli nascenti nel suo interno quando vengono attraversati da una onda di densità e comincia a formarsi un braccio a spirale. L'onda può accelerare i vari processi che ora descriverò.

Le condizioni nei circa 4000 complessi molecolari giganti che si trovano entro 13 000 parsec dal centro della Galassia sembrano fatte su misura per la formazione di nuove stelle. Tanto per cominciare, ogni complesso è ricco di materia. La massa di un complesso tipico raggiunge parecchie centinaia di migliaia di masse solari e il suo diametro è di circa 50 parsec. È l'oggetto più pesante della Galassia (a meno che nel centro della Galassia non esista un buco nero supermassiccio). Si pensa che la massa di un complesso molecolare gigante sia quasi interamente costituita da molecole di idrogeno; a una temperatura di 20 kelvin, la nube è troppo fredda perché le molecole possano dissociarsi in atomi. Inoltre la nube è in molti punti troppo fredda perché le molecole di idrogeno possano emettere quantità rivelabili di radiazione alle loro caratteristiche lunghezze d'onda. Si deve quindi dedurre la presenza di idrogeno molecolare. La nube viene meglio rivelata dall'emissione (a 2,6 millimetri) della molecola immediatamente meno comune, il monossido di carbonio, e dalle emissioni di molecole ancor meno comuni quale la formaldeide.

Ai fini della formazione delle stelle è importante anche un altro componente di un complesso molecolare, la polvere. Le particelle di polvere sono punti nei quali si può raccogliere il gas circostante. Anche la polvere scherma le stelle nascenti dalla radiazione ultravioletta, che impedirebbe la condensazione. Si ritiene che per ogni 100-200 grammi di idrogeno molecolare del complesso molecolare gigante vi sia un grammo di polvere.

Due complessi vicini di monossido di carbonio e di polvere (e forse di idrogeno in abbondanza) sono stati studiati particolarmente a fondo. Si tratta del complesso di Orione (che ha il centro nella luminosa regione di idrogeno ionizzato e di stelle giganti appena nate chiamata la Grande Nebulosa di Orione) e il complesso dell'Ofiuco (che blocca la luce proveniente dal centro della Galassia). Nel complesso di Orione, Becklin, Neugebauer e i loro collaboratori del Cal Tech, insieme a Frank J. Low dell'Università dell'Arizona, hanno trovato prove alle lunghezze d'onda infrarosse della presenza di oggetti condensati che paiono essere intrinsecamente molto rossi. Ognuno di essi potrebbe essere una stella appena nata ancora immersa in uno spesso bozzolo di polvere che la radiazione ultravioletta della stella non ha ancora spazzato via completamente. La radiazione riscalda la polvere che successivamente irraggia nell'infrarosso. La stella può essersi formata in primo luogo quando il raffreddamento di una piccola parte del complesso ha fatto diminuire la pressione causata dal riscaldamento del gas. La regione raffreddata avrebbe in tal modo cominciato a collassare sotto l'effetto della sua stessa gravitazione.

Nel complesso di Ofiuco è stato scoperto un gruppo di 30 stelle da Gary L. Grasdalen dell'Università del Wyoming, Stephen E. e Karen M. Strom dell'Università dell'Arizona e Frederick J. Vrba dell'US Naval Observatory. Le stelle non sono state rivelate prima perché sono nascoste da almeno 30 magnitudini di assorbimento ottico; fu necessaria una ricerca nell'infrarosso per scoprirle. Osservazioni più recenti di Charles J. Lada e Bruce A. Wilking dell'Università dell'Arizona in una nube di polvere molto densa nella stella Rho Ophiuchi hanno rivelato la presenza di 20 stelle simili la cui radiazione ottica è attenuata fino a 100 magnitudini. Le stelle distano solo mezzo parsec da quelle scoperte in precedenza.

L'evidenza oggi disponibile fa ritenere che i complessi molecolari giganti originano spontaneamente stelle con masse non superiori a poche masse solari. In particolare sembra che essi diano origine a stelle delle classi spettrali *B*, *A*, *F* e *G*. (Il Sole è una stella della classe *G*.) George H. Herbig e i suoi collaboratori del Lick Observatory hanno scoperto nei pressi dei bordi dei ben noti complessi molecolari giganti alcuni gruppi di piccole, fioche e nebulose stelle che si dice siano giovani. Quasi tutte hanno una luminosità variabile in modo irregolare. Esse sono chiamate stelle T Tauri e, forse, sono state prodotte dall'ininterrotta condensazione di noduli e possono essere fuoriuscite dal complesso. Vengono spesso osservate in luoghi nei quali la radiazione ultravioletta di stelle pesanti appena nate ha spazzato via il gas e la polvere della nebulosa oscura.

Pare che la formazione di stelle di un altro tipo sia tipica di luoghi nei quali una nebulosa di emissione con stelle supergi-

ganti *O* e *B* immerse in essa si trova vicina a un complesso molecolare gigante. La più chiara esposizione di ciò che accade in tali casi è quella di Bruce G. Elmegreen della Columbia University e di Lada. Secondo Elmegreen e Lada, le stelle *O* e *B* emettono radiazione ultravioletta la cui pressione fa accumulare gas freddo e polvere sul bordo esterno del complesso. Il risultato è la condensazione, in questa regione, di protostelle. Sembra che in luoghi quali la nebulosa di Orione il processo si sviluppi per fasi. In questa nebulosa un gruppo di stelle supergiganti *O* e *B* si sta attenuando dopo una vita media di qualche decina di milioni di anni. La radiazione proveniente da queste stelle ha innescato la formazione di una più giovane generazione di stelle supergiganti *O* e *B* che emettono oggi a loro volta radiazione che si apre la strada lentamente, ma persistentemente, nel complesso molecolare gigante, dove si formerà presumibilmente un terzo gruppo di stelle supergiganti *O* e *B*. Non è ancora ben chiaro perché il processo possa dare origine a giganti e supergiganti *O* e *B* anziché alle più piccole stelle *B*, *A*, *F* e *G* che condensano spontaneamente nel complesso.

La formazione di stelle nei globuli

Mi occupo ora della classe di nubi oscure chiamate globuli, di cui ne sono stati scoperti circa 200 entro una distanza di 500 parsec dal Sole. Essi hanno proprietà notevolmente simili. Ciascuno di essi è oscuro e distinto e su una lastra fotografica è quasi circolare. Non vi è dubbio che siano sferici. Il loro raggio varia da 0,2 a 0,6 parsec, la loro massa da 20 a 200 masse solari e la loro temperatura interna da cinque a quindici kelvin. I globuli sono impenetrabili alla luce visibile. D'altra parte, alcune immagini registrate alle lunghezze d'onda dell'infrarosso vicino sia fotograficamente sia con tecniche elettroniche di formazione di immagine mostrano le stelle dietro ai globuli. L'attenuazione della radiazione infrarossa da parte di tali stelle consente una stima del contenuto di polvere del globulo.

Le osservazioni radio a 2,6 millimetri mostrano che i globuli sono ricchi di monossido di carbonio. Ora sono state individuate altre molecole, in particolare formaldeide e ammoniaca. È quindi evidente che un globulo è una piccola e spesso isolata sferetta di oscurità del tutto simile in composizione a un complesso molecolare gigante. È probabile che il globulo sia composto principalmente da idrogeno molecolare troppo freddo per emettere radiazioni rivelabili.

I dati radio hanno anche mostrato che parecchi globuli stanno collassando sotto l'azione della loro stessa gravitazione. La velocità di collasso è di circa mezzo chilometro al secondo, ossia mezzo parsec per milione di anni. Dal momento che il raggio di un globulo tipico è di mezzo parsec, un milione di anni è all'incirca il tempo necessario per il completamento del collasso. Ancora una volta il collasso è un punto critico dei modelli di formazione delle stelle. In quasi tutti i modelli il centro del globulo collassa più rapidamente della periferia e quindi si forma un nodulo. Il collasso trasforma in energia cinetica l'energia potenziale gravitazionale della materia incidente. Alla fine l'energia nel centro fa aumentare la temperatura della materia tanto da dare inizio alla fusione termonucleare. Ciò indica la nascita di una stella. Se la stella è grande, essa emette radiazioni sufficienti a spazzar via il gas e la polvere che la circondano.

In breve, un globulo dovrebbe dare origine a una sola stella in circa un milione di anni. Considerando la scala dei tempi del processo e valutando il numero di globuli della Via Lattea, si conclude che i globuli potrebbero spiegare la formazione di 25 000 stelle per ogni milione di anni, ossia circa un sesto del tasso complessivo con cui si formano le stelle nella Galassia. (La formazione di stelle in complessi molecolari giganti è forse un processo più fecondo.) Tuttavia, nel 1977 Richard Schwartz dell'Università del Missouri osservò ciò che pare essere una coppia di stelle nebulose nascenti espulse da un globulo molto scuro a circa 300 parsec dal sistema solare nella regione meridionale del cielo contraddistinta dalla costellazione della Vela. La creazione di due stelle anziché di una è sorprendente e altrettanto lo è la loro espulsione dal globulo. Ci si domanda da dove venga l'energia necessaria per espellerle. Le due stelle sono collegate da una striscia luminosa che potrebbe essere descritta con una certa fantasia come un cordone ombelicale stellare.

Mi piacerebbe dare un ultimo esempio per suggerire quanto resta da scoprire. I ricercatori che stanno tentando di realizzare modelli dei processi di formazione delle stelle hanno tacitamente supposto che sia necessaria la protezione di una nube di polvere. All'interno di tale nube la temperatura è bassa e l'accumulo di materia avviene indisturbato: la polvere scherma l'interno della nube dalla distruzione da parte di radiazione ultravioletta proveniente da sorgenti esterne. Si considerino, però, le Nubi di Magellano che, trovandosi nella corona galattica, negano la generalizzazione che tutto il contenuto della corona sia vecchio.

Nella Grande Nube di Magellano esiste un raggruppamento di circa 50 stelle luminose *O* e *B* note come Costellazione I di Shapley. È probabile che la loro età non sia superiore a 20 milioni di anni. Le loro velocità sono dell'ordine di soli 10 chilometri al secondo; perciò ogni stella si è spostata di non più di 200 parsec dal luogo in cui sono ora le stelle, le misurazioni alla lunghezza d'onda di 21 centimetri hanno rivelato cinque milioni di masse solari di idrogeno atomico neutro, mentre le misure a lunghezza d'onda ottiche hanno rivelato 60 000 masse solari di idrogeno ionizzato. Ma il cielo in tale regione è trasparente. Probabilmente là non c'è alcun complesso molecolare gigante; c'è poco o niente idrogeno molecolare e non c'è polvere cosmica. Come si sono formate le stelle in tale regione?

Impollinazione sottomarina

I pochi generi di piante che fioriscono sotto la superficie del mare devono essere in grado di diffondere e catturare il polline in un ambiente subacqueo. Quali adattamenti consentono loro di riprodursi?

di John Pettitt, Sophie Ducker e Bruce Knox

Nel ciclo vitale di una angiosperma il momento dell'impollinazione è fra i più importanti. Si sono pertanto evolute elaborate strategie per trasferire il polline da un fiore all'altro: alcune piante fanno affidamento sul vento, altre sulle api, sui pipistrelli o su altri animali. In alcuni gruppi di angiosperme, tuttavia, questi sistemi non funzionano in quanto i fiori sbocciano sott'acqua. Le piante di questo tipo, che colonizzano gli ambienti costieri nelle regioni tropicali e temperate, appartengono tutte alle angiosperme monocotiledoni.

Come è possibile l'impollinazione subacquea? A prima vista la risposta è molto semplice: il polline viene distribuito dal movimento dell'acqua e in particolare dall'azione delle onde. Il metodo sembra perciò essere in qualche modo analogo a quello dell'impollinazione effettuata dal vento, ma, a un esame più attento, c'è una interessante complicazione. Fra le angiosperme terrestri un elemento chiave per assicurare il successo della riproduzione è il «riconoscimento» chimico fra le molecole di proteina o glicoproteina (che ha catene laterali di zucchero) situate sul granulo di polline o sul tubetto pollinico e le molecole di glicoproteina presenti sulla superficie recettiva dello stimma o delle cellule dello stilo del pistillo, l'organo femminile del fiore. Queste molecole sono solubili in acqua, una caratteristica che è difficilmente compatibile con il riconoscimento del polline sott'acqua. Viene quindi spontaneo chiedersi se le piante marine hanno sviluppato un tipo di riconoscimento modificato? È stato questo l'argomento del lavoro svolto dal nostro gruppo.

Innanzitutto va premesso che le monocotiledoni marine non solo non vanno confuse con le alghe, ma hanno anche scarse affinità con le monocotiledoni terrestri. Esse comprendono 12 generi per un totale di circa 50 specie; a esse appartengono le famiglie delle *Potamogetonaceae* e delle *Hydrocharitaceae*. La *Zostera marina* e il genere *Phyllospadix* sono esempi familiari nelle regioni temperate. I generi *Halophila* e *Thalassia* sono ampiamente distribuiti nei tropici.

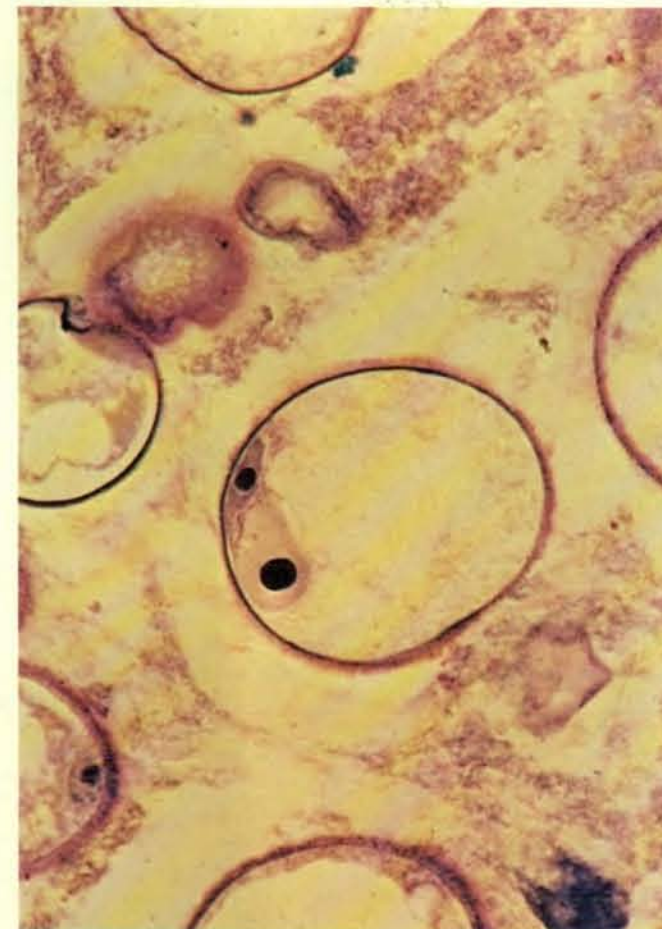
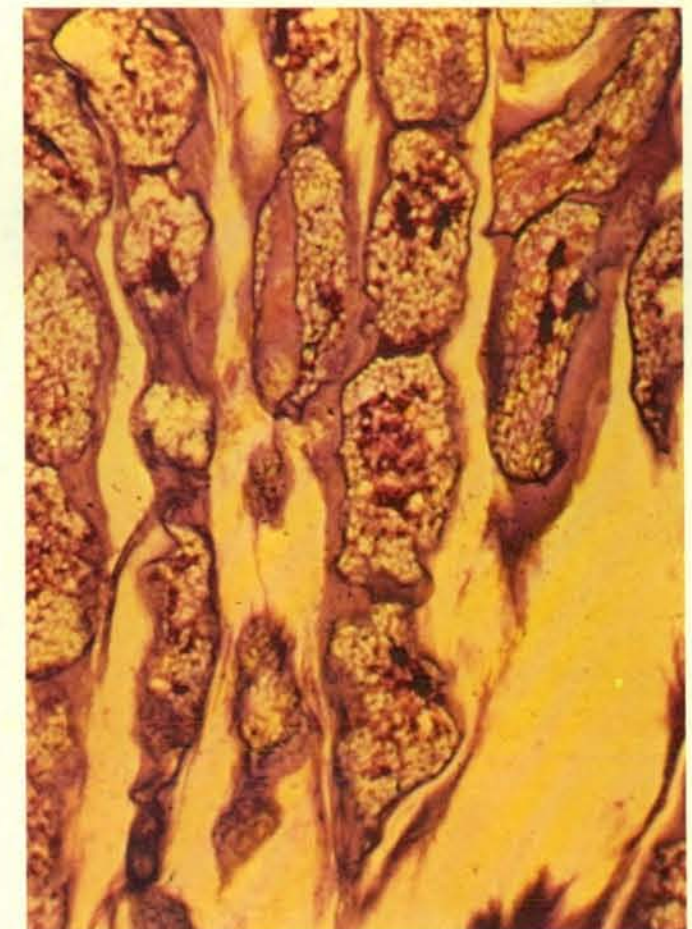
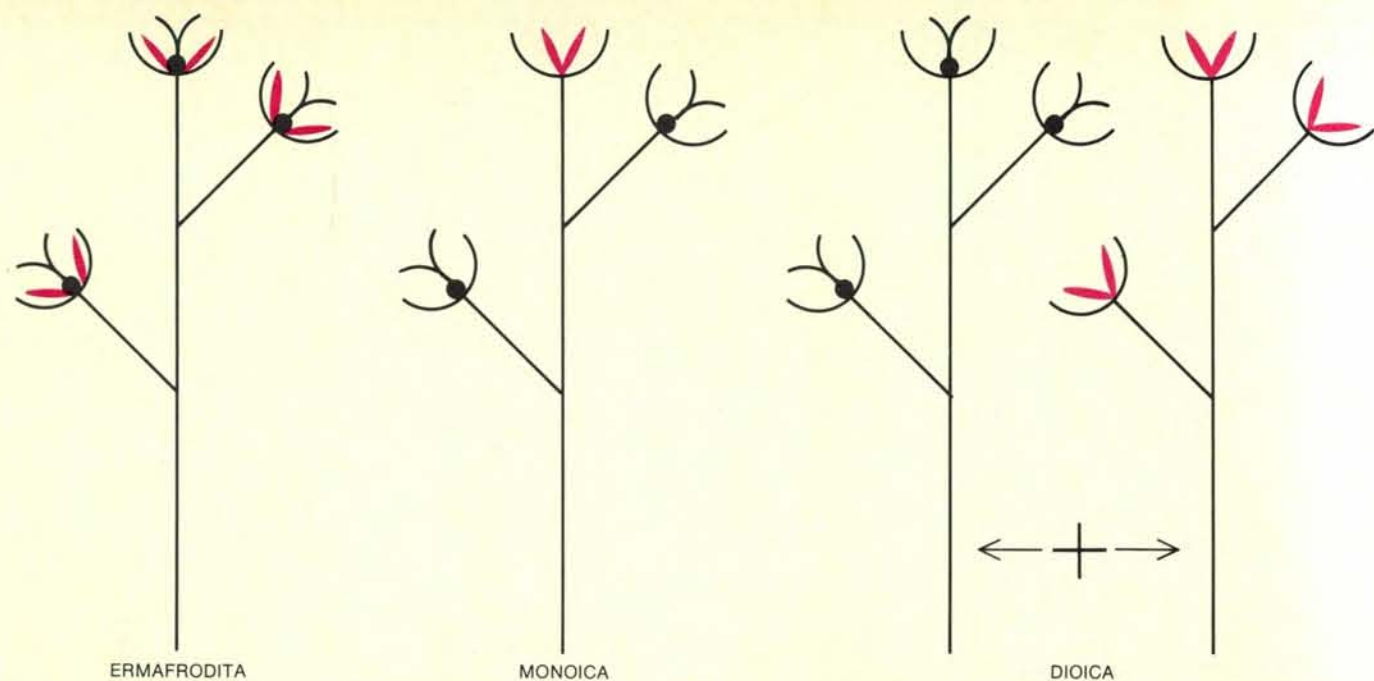
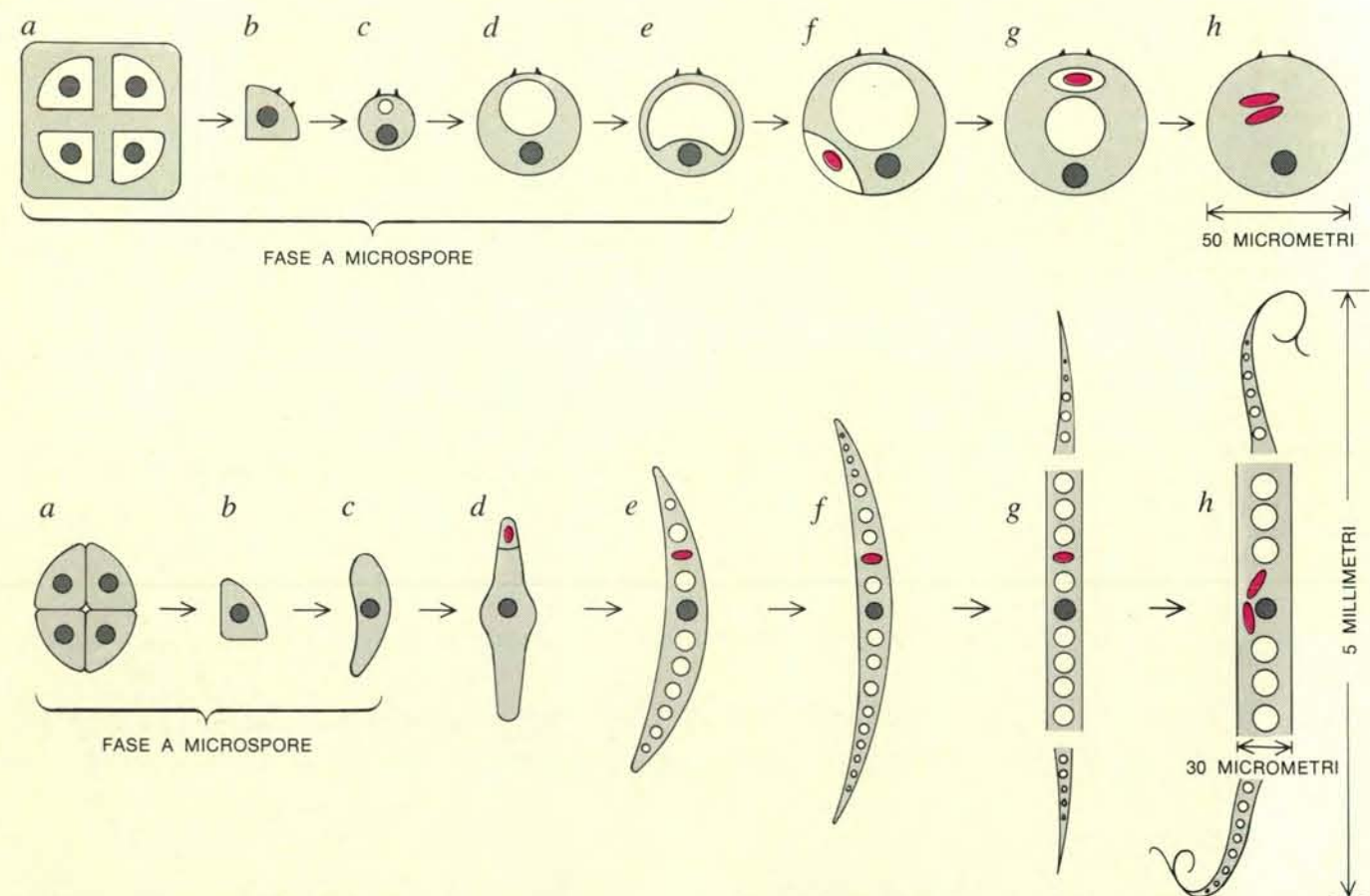
Una vasta ricerca condotta da Cornelis den Hartog dell'Università di Nimega ha dimostrato che la maggior concentrazione delle monocotiledoni marine si trova nella regione indomalese dove si sono trovati 8 dei 12 generi esistenti. Un'altra concentrazione è stata rinvenuta nell'area tropicale del Nuovo Mondo e, in particolare, nel Mar dei Caraibi. La distribuzione fa sorgere molti interrogativi sulle origini geografiche, sulle migrazioni e sulla speciazione delle piante. Per esempio, il genere *Thalassia* comprende due specie morfologicamente simili che sembrano imparentate molto da vicino e che possono derivare da una comune popolazione ancestrale. Una delle specie si trova nell'oceano Indiano e nel Pacifico occidentale e l'altra nel Mar dei Caraibi e nel golfo del Messico. Le specie non si riproducono fra loro. Questo avviene perché sono separate geograficamente o perché non hanno affinità genetica? Se venissero in contatto tra loro, si riprodurrebbero con successo o mostrerebbero un isolamento riproduttivo? La domanda non ha ancora avuto risposta per *Thalassia*, ma l'ha avuta per altre monocotiledoni marine del genere *Amphibolis*. Anche questo genere comprende due specie, ma non c'è grande distanza che le separa. Crescono insieme lungo le coste meridionali e occidentali dell'Australia. Sembra esserci tuttavia una barriera biologica all'incrocio interspecifico. Sebbene i periodi di fioritura delle due specie possano coincidere o almeno sovrapporsi per qualche tempo, non si sono osservati discendenti ibridi.

Per quanto riguarda la riproduzione, tutte le angiosperme si dividono generalmente in una o nell'altra delle seguenti tre categorie. La maggior parte è ermafrodita, cioè ogni fiore comprende sia i pistilli femminili sia le antere maschili che liberano il polline. Alcune piante, tuttavia, sono monoiche (dal greco «una sola casa»): un dato fiore ha solo pistilli o solo antere, ma una singola pianta porta entrambi i tipi di fiore. Solo il 3 o 4 per cento di tutte le angiosperme sono dioiche («con due case»); tutti i fiori di alcune piante hanno solo pistilli e tutti i fiori di altre piante hanno solo antere. La frequenza di questi stati riproduttivi nelle angiosperme marine differisce notevolmente da quella delle forme terrestri. Ben nove dei dodici generi marini sono dioici.

Per quanto riguarda la sessualità del fiore, la sequenza degli avvenimenti nella riproduzione sessuale è fondamentalmente la stessa in tutte le angiosperme. Quando il fiore matura, l'antera si apre per liberare i granuli di polline; ogni granulo contiene i gameti maschili. I granuli devono allora essere portati su un pistillo maturo e aderire allo stimma recettivo del pistillo. Se il granulo viene riconosciuto come compatibile, germina e incomincia a produrre una struttura tubulare. Questo tubetto pollinico penetra nella cuticola dello stimma, cresce attraverso lo stilo del pistillo e alla fine raggiunge l'ovario, dove sono contenuti i gameti femminili. I gameti maschili passano attraverso il tubetto pollinico nell'ovario dove avviene la fecondazione e a tempo debito maturano i semi. Nel caso che il granulo di polline non sia compatibile, o non si forma il tubetto pollinico, o non riesce a penetrare nello stimma, o non riesce a crescere attraverso lo stilo. In ognuno dei tre casi non c'è unione fra i gameti maschile e femminile.

Le angiosperme dominano la vegetazione terrestre e le forme acquatiche sono derivate senz'altro da esse in tempi relativamente recenti. Tuttavia queste ultime presentano molte differenze morfologiche e strutturali rispetto alle angiosperme terrestri e alcune delle differenze hanno certamente valore di adattamento all'ambiente sottomarino. Tuttavia, le forme marine possiedono ancora caratteristiche della loro eredità terrestre e la sequenza di base degli avvenimenti della loro riproduzione non è apparentemente differente da quella delle angiosperme terrestri.

Gli adattamenti fiorali delle angiosperme possono assumere una varietà di forme, ma tutti devono permettere

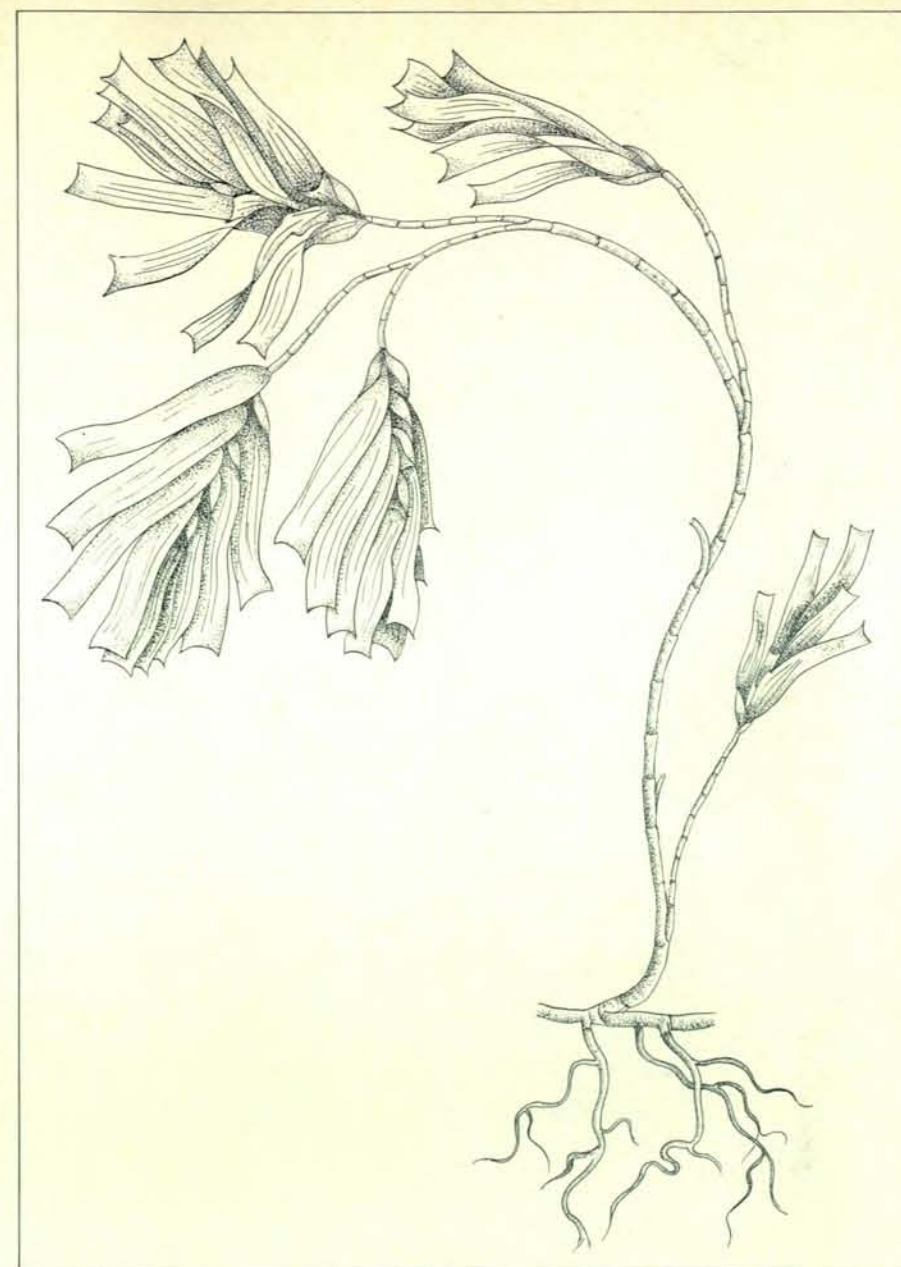
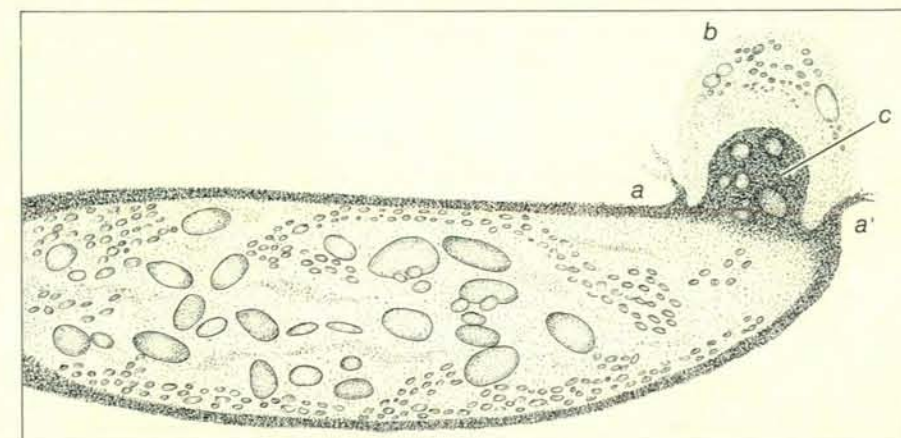
Questa colonia di *Thalassia*, uno dei 12 generi di monocotiledoni marine che vengono impollinate sott'acqua, cresce su una piattaforma battuta dalle onde lungo le coste del Kenya. La *Thalassia* è visibile in questo caso per le condizioni di bassa marea; durante l'alta marea è sommersa.

Questa fotografia al microscopio permette il confronto fra due tipi diversi di polline di piante marine. A sinistra sono riprodotti i granuli immaturi del polline di *Thalassia*, visti in una sezione dell'antera, l'organo maschile della pianta. Come appare evidente nella parte centrale dell'immagine, i granuli sono allo stadio di sviluppo a due cellule; hanno incominciato a formarsi masserelle di sostanza viscosa *(che si è colorata in rosso)* che circonderà i granuli di polline quando saranno maturi. A destra sono visibili alcuni filamenti lineari del polline di *Halophila*; ogni filamento è un tubo mucillaginoso che contiene quattro granuli di polline. I tubi sono riconoscibili per la colorazione rossa.

Le angiosperme presentano i loro organi sessuali in una di queste configurazioni. La maggior parte è ermafrodita; ogni fiore ha antere maschili *(in colore)* che producono il polline e pistilli femminili *(in nero)* che catturano il polline liberato. Alcune sono monoiche: i fiori maschili *(in colore)* e femminili *(in nero)* crescono sulla stessa pianta, come al centro. Alcune angiosperme, forse il 4 per cento del totale, sono dioiche; ogni pianta porta solo fiori femminili o fiori maschili, come nell'esempio a destra. Dei 12 generi di angiosperme marine, il 75 per cento è dioico.

Confronto tra lo sviluppo del polline in un'angiosperma terrestre *(in alto)* e in una marina *(in basso)*. La produzione dei granuli di polline del loglio incomincia con una tetrade di microspore. Dopo la liberazione dalla tetrade ogni microspora si ingrandisce; la fase a microspore termina cinque stadi più tardi, una divisione *(f)* forma una cellula vegetativa *(in nero)* e una cellula madre dello sperma *(in colore)*. Una seconda divisione *(h)* produce le due cellule spermatiche del granulo maturo; la maturazione impiega due settimane e il granulo raggiunge i 50 micrometri di diametro. Anche la produzione del granulo filamentoso del polline di *Amphibolis* inizia con una tetrade di microspore. Tre stadi dopo, la cellula madre dello sperma *(in colore)* giace a una estremità del granulo allungato *(d)* e occupa una posizione centrale prima di dividersi in due cellule spermatiche *(h)*. La maturazione dura 16 settimane; il granulo arriva a 30 micrometri di diametro e 5 millimetri di lunghezza.

un'impollinazione efficace per la regolazione del livello di esoincrocio, per il nutrimento dell'embrione e per la dispersione dei semi in quantità sufficiente al mantenimento della specie. Il primo di questi provvedimenti, l'impollinazione efficace, ha portato allo sviluppo di fiori grandi (in quelle piante in cui l'impollinazione dipende da visitatori animali che ne vengono attratti). Le piante impollinate dal vento, d'altra parte, non hanno bisogno di offrire tali attrazioni e quindi le loro strutture fiorali ancillari ne sono completamente o quasi prive.

Lo stesso criterio è valido per le angiosperme marine. In queste piante impollinate dall'acqua, le strutture fiorali necessarie a una pianta che viene impollinata dagli insetti (i petali della corolla o i sepali del calice colorati) sono poche e, se sono presenti, offrono poca o nessuna mostra di sé. La *Thalassia* è un'eccezione, i suoi fiori portano strutture simili a petali colorati, che sono rosso brillante quando il fiore si apre. Qualunque sia la funzione di queste strutture appariscenti, non può essere quella di attirare l'attenzione di un potenziale animale impollinatore.

Che il fiore attragga o meno, due dei suoi compiti vitali sono la produzione e la recezione del polline. Nei fiori delle piante marine, gli organi adibiti a queste funzioni sono ben sviluppati. Gli stami maturi liberano granuli di polline galleggianti e i pistilli hanno stimmi recettivi per catturare i granuli portati dalla corrente.

Sempre per garantire il successo della riproduzione, e in particolare la regolazione dell'esoincrocio, è vantaggioso il fatto che la maggior parte dei generi delle piante marine siano dioici. La segregazione delle funzioni sessuali fra individui differenti è una strategia di esoincrocio molto efficace: impone il movimento del polline fra le piante e così produce una progenie con la più alta frequenza possibile di variabilità genetica. Tuttavia le piante dioiche non offrono un metodo di riproduzione molto efficace. La ragione è che solo metà delle piante di una popolazione, le femmine, contribuiscono alla produzione dei semi, un fatto che influisce sull'ultima fase della riproduzione. L'incidenza notevolmente alta di piante dioiche fra le piante marine, significa presumibilmente che i vantaggi nei confronti della sopravvivenza attribuibili all'esoincrocio superino gli svantaggi della diminuita produzione di semi.

La fioritura nelle piante terrestri avviene in risposta ai cambiamenti della durata dell'insolazione, della temperatura o di altri fattori. Quali sono i fattori che stimolano l'inizio e lo sviluppo dei fiori nelle piante marine? All'inizio del secolo, il botanico svedese Nils Svedelius descrisse la fioritura del genere *Enhalus*, una monocotiledone dell'oceano Indo-Pacifico, che egli studiò a Ceylon. Egli osservò una cosa che era nota da tempo ai pescatori del luogo: la fioritura di *Enhalus* è sincronizzata con il ciclo delle maree. Se le ma-

Questa pianta marina del genere *Amphibolis* fu usata dagli autori in Australia negli esperimenti di impollinazione. La pianta, a fusti eretti, raggiunge un'altezza variabile tra i 20 e gli 80 centimetri.

In questo disegno basato su una microfotografia, si può osservare come l'apertura di un granulo di polline di *Amphibolis*, avvenuta dopo che il granulo si è attaccato allo stimma, sia dovuta alla frammentazione della parete del polline *(a, a')*, disciolta dall'azione di un enzima. Dall'apertura esce una bolla mucillaginosa *(b)*, e una regione di denso citoplasma *(c)* preannuncia il tubetto pollinico.

ree più alte del ciclo - le maree sigiziali che seguono la luna nuova e la luna piena - si verificavano durante il giorno, la fioritura avveniva in coincidenza.

Quattro anni fa scoprimmo una correlazione simile nel comportamento riproduttivo di *Thalassia*. Le colonie che studiammo crescono su una piattaforma di acqua poco profonda lungo la costa del Kenya. Durante le basse maree, quando la variazione del livello di acqua dovuto alla marea è al minimo, i fiori femminili e maschili incominciano a emergere dalle foglie che li racchiudono. Con l'inizio del successivo periodo di marea sigiziale, le piante sono in piena fioritura e il polline viene liberato. In questo modo la *Thalassia* fiorisce quando la circolazione dell'acqua sulla piattaforma è al massimo.

Ogni fiore incomincia ad appassire alcune ore dopo l'apertura. Il periodo complessivo di fioritura per tutte le piante della popolazione è solo di pochi giorni. Quando noi raccogliemmo le piante alla fine delle maree sigiziali, stavano o producendo frutti o incominciando a sviluppare fiori che sarebbero maturati alle successive maree sigiziali. La liberazione del seme avveniva anch'essa in sincronia con il ciclo delle maree. Per una popolazione di piante che cresce in un ambiente limite questo è un mezzo estremamente efficace per favorire la distribuzione del polline e la dispersione dei semi.

In altre piante marine la fioritura sembra essere regolata dalle stagioni piuttosto che dalle maree. Per esempio, nelle colonie di *Amphibolis* lungo la costa dello stato di Victoria in Australia, le prime strutture fiorali appaiono all'inizio dell'inverno australe (maggio) e i fiori generalmente raggiungono la maturità nel periodo che va dall'inizio dell'estate a metà estate (da metà novembre alla fine di dicembre). I semi di *Amphibolis* germinano quando il frutto è ancora attaccato alla pianta madre. I frutti maturi con i loro germogli vengono diffusi durante i mesi invernali, da maggio ad agosto. Il periodo di fioritura delle colonie di *Amphibolis* dell'Australia occidentale è in qualche modo in ritardo rispetto a quello delle piante di Victoria, fatto che suggerisce che l'inizio della formazione delle strutture fiorali e la loro maturazione sono regolati in base alla lunghezza del giorno.

Il polline delle monocotiledoni marine ha attirato l'interesse dei naturalisti almeno dalla fine del diciottesimo secolo. Fu allora che Filippo Cavolini descrisse e disegnò a Napoli il polline di *Posidonia*. Cavolini osservò che con l'apertura delle antere appariva nell'acqua una nuvola di filamenti. Alcuni decenni più tardi, nel 1826, il botanico francese Charles Gaudichaud-Beaupré osservò e disegnò il polline similmente filiforme di *Amphibolis* in esemplari raccolti nella Shark Bay nell'Australia occidentale. Alcuni decenni dopo, Edouard Bornet, un altro botanico francese, descrisse lo sviluppo del polline in colonie dell'esemplare marino *Cymodocea* nelle acque del Mediterraneo.

All'inizio del ventesimo secolo il botanico svedese Otto Rosenberg fissò, sezionò e colorò una serie di antere di *Zostera*. Fu in grado di tracciare la differenziazione del polline dalla sua origine alla meiosi (il tipo di divisione cellulare che dimezza il numero di cromosomi), fino alla maturità. In questo genere il polline è filiforme. Il processo di differenziazione impiega molte settimane e il filamento del polline maturo raggiunge una lunghezza di circa tre millimetri. I gameti maschili portati dal polline hanno un diametro che non supera i 5-10 micrometri.

Negli ottant'anni successivi al lavoro di Rosenberg sul polline di *Zostera marina*, ricerche simili hanno rivelato le caratteristiche generali dello sviluppo del polline in altre piante marine. Attraverso le metodiche per la ricerca di componenti cellulari specifici è stato possibile non solo risolvere la struttura e la composizione dei granuli di polline delle monocotiledoni marine con notevole precisione, ma anche dimostrare il ruolo di questi componenti nella riproduzione.

I granuli pollinici delle monocotiledoni marine hanno determinate caratteristiche che sembrano essere variazioni adattative derivate dal polline degli antenati terrestri. Per apprezzare l'estensione di questi cambiamenti adattativi è utile in primo luogo considerare come avviene l'impollinazione sulla terraferma. Il granulo di polline di una angiosperma terrestre è generalmente sferico e ha un diametro di 30-60 micrometri, quando è liberato dalle antere. Secondo le specie di pianta, il granulo maturo è formato da due o tre cellule. Se ci sono solo due cellule, una è vegetativa, e l'altra è il gamete progenitore che si divide per formare due cellule spermatiche dopo la germinazione del polline. In un granulo di polline con tre cellule, una è ancora vegetativa e le altre due sono cellule spermatiche.

La parete del granulo di polline che circonda le cellule spermatiche è formata da due strati che differiscono in origine, struttura e composizione chimica. Lo strato più esterno è formato da sporopollinina, una sostanza dura che resiste alla rottura ed è spesso ornata da elaborate sculture. Lo strato interno è privo di sculture ed è formato da sottili fibrille di cellulosa trattenute in una matrice di polisaccaridi. Il materiale per la costruzione dello strato più esterno si origina nel tappeto, il tessuto nutritivo dell'antera; lo strato più interno è un prodotto del protoplasma del granulo di polline.

Nella parete a due strati del granulo di polline sono generalmente presenti aperture o pori; variano notevolmente in forma, numero e posizione secondo la specie. Quando il polline germina, il tubetto pollinico emerge attraverso una delle aperture. Alcuni enzimi, compresa la fosfatasi acida, sono incorporati nello strato interno del granulo; la loro concentrazione è particolarmente alta vicino alle aperture. Nei granuli di polline che non hanno aperture, il tubetto pollinico può emergere attraverso la parete del granulo in qualunque punto; gli enzimi sono distribuiti uniformemente intorno alla circonferenza del granulo, nello strato interno. In entrambi i casi il risultato è che gli enzimi sono portati all'estremità del tubetto pollinico quando questo emerge. Gli enzimi sono anche trasferiti allo strato esterno della parete del granulo dalle cellule del tappeto dell'antera in una fase avanzata dello sviluppo del polline; questo trasferimento fu dimostrato per la prima volta all'inizio degli anni settanta da John Helsop-Harrison, che era allora ai Royal Botanical Gardens a Kew, e da uno di noi (Knox). Fra gli enzimi trasferiti c'è l'esterasi, che viene ora ritenuta la proteina caratteristica dello strato esterno.

I granuli di polline delle piante marine sono di tre tipi fondamentali: filiformi, sferici ed ellissoidali. I granuli filiformi sono caratteristici della *Zostera marina* e dei generi *Posidonia, Amphibolis, Heterozostera* e *Thalassodendron*. Questi corpi filiformi che sono lunghi fino a 5 millimetri e hanno un diametro che va da 10 a 30 micrometri, sono morfologicamente diversi dai pollini delle piante terrestri. Il granulo è privo dello strato esterno di sporopollinina; la parete del polline è formata solo da microfibrille di cellulosa immerse in una matrice polisaccaridica. Quando granuli di polline filiforme freschi, maturi vengono sottoposti al saggio della fosfatasi acida (l'enzima trovato nello strato interno delle pareti del granulo di polline delle angiosperme terrestri), la reazione è positiva.

La superficie dei granuli di polline filiformi è ricoperta da una sottile pellicola di materiale che comprende carboidrati e proteine. Entro la pellicola c'è l'enzima esterasi (il prodotto delle cellule del tappeto trovato nella parete esterna del polline delle piante terrestri); esso viene depositato nella fase finale della maturazione del polline. Così, benché vi siano differenze morfologiche fra il polline filiforme delle piante marine e il polline delle piante terrestri, vi sono tuttavia somiglianze strutturali e biochimiche.

Due monocotiledoni marine tropicali e subtropicali, *Thalassia* ed *Enhalus*, liberano granuli di polline sferici. Quelli di *Thalassia* hanno un diametro di 100 micrometri e quelli di *Enhalus* di 150 micrometri; entrambi sono privi di aperture. Il polline delle due piante assomiglia molto da vicino al polline delle angiosperme terrestri, avendo due strati distinti, esterno e interno. Gli enzimi dello strato interno comprendono la fosfatasi acida; nella sottile pellicola che copre lo strato esterno è presente la esterasi.

Nel corso degli studi nell'oceano Indiano notammo che i granuli di polline di *Thalassia* venivano liberati nel mare cementati in una sostanza viscosa viscida, insolubile; la massa di polline assomigliava alle uova degli anfibi, sebbene il polli-

Gli stimmi di *Amphibolis*, le strutture ramificate visibili in questa fotografia eseguita al microscopio elettronico a scansione, hanno catturato molti granuli filamentosi di polline liberati dalla pianta maschile e inviati a fiori femminili dagli sperimentatori. La struttura è ingrandita 20 volte.

ne sia molto più piccolo. Carl Julius Fritsche, un chimico tedesco del diciannovesimo secolo, portò l'attenzione sulla presenza di questa sostanza viscosa nelle piante acquatiche. Prendemmo in considerazione l'ipotesi che la sostanza insolubile potesse giocare un ruolo nella impollinazione di *Thalassia*, forse attaccando i granuli di polline agli stimmi recettivi. Mentre seguivamo lo sviluppo del polline in *Thalassia*, individuammo il sito in cui si originava la sostanza viscosa. Il protoplasma delle cellule del tappeto invade la cavità dell'antera ed è modificato nella sostanza viscosa durante lo stadio finale dello sviluppo del polline.

Per determinare la composizione della sostanza viscosa la separammo dal polline mediante centrifugazione. Le analisi dimostrarono che la sostanza viscosa è composta per il 5 per cento da carboidrati. Mediante cromatografia riuscimmo a identificare gli zuccheri contenuti nella frazione di carboidrati. Il 50 per cento è rappresentato da mannosio; il resto è xilosio, galattosio, glucosio, arabinosio e ramnosio. La tecnica dell'elettroforesi fu impiegata per cercare le proteine e le glicoproteine della sostanza viscosa in base alle loro dimensioni molecolari. Fu trovato un singolo componente di glicoproteina.

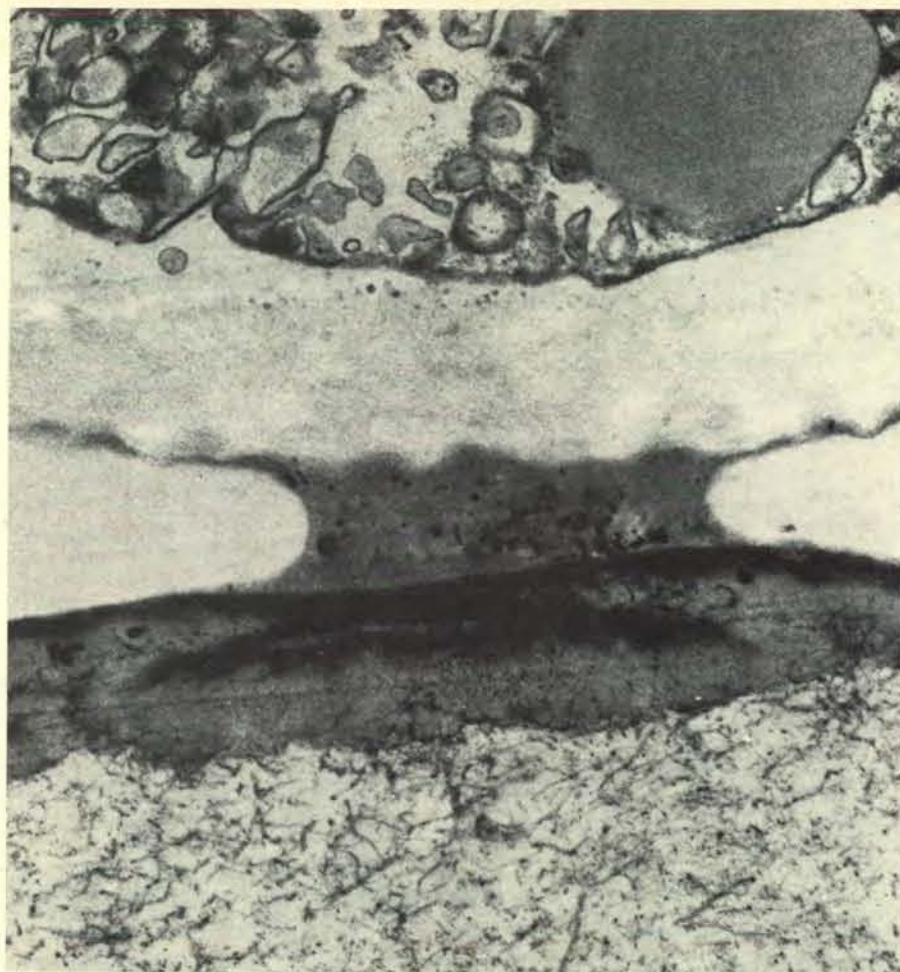
L'osservazione al microscopio del polline centrifugato dimostrò che anche dopo la centrifugazione, un sottile strato di pellicola avvolgeva ciascun granulo. Queste scoperte suggeriscono che anche dopo che il moto ondoso ha frantumato la massa del polline, un po' di sostanza viscosa aderisce ancora ai singoli granuli di polline per servire come adesivo alla superficie dello stimma.

La terza forma di polline delle monocotiledoni marine è prodotta dalle piante del genere *Halophila*. Questi granuli ellissoidali misurano circa 40 × 80 micrometri e sono privi di aperture. Il polline viene liberato dalle antere in filamenti; ogni filamento è un tubo mucillaginoso che contiene quattro granuli di polline. I quattro granuli sono cellule «sorelle» che si formano dalla stessa cellula madre durante la meiosi. La parete cellulare è priva di strato esterno e mostra la tipica struttura di uno strato interno, formata da microfibrille in una matrice polisaccaridica; è presente la fosfatasi acida. Il tubo mucillaginoso ha struttura simile ed è ricoperto da una sottile pellicola di proteine, compresa l'esterasi, derivate dal tappeto. La presenza di una pellicola superficiale suggerisce un ruolo nel riconoscimento e nell'attacco del polline.

La cattura e il trattenimento dei granuli di polline alla deriva dipende anche dalla superficie recettiva dello stimma della monocotiledone marina. In alcuni generi la superficie è liscia o costolata; in altri è coperta di piccole prominenze. Entrambi i tipi di superficie sono dotati di una sottile cuticola rivestita a sua volta da una secrezione superficiale.

Le secrezioni dello stimma sono direttamente paragonabili a quelle di certe piante terrestri. Uno di noi (Knox), in collaborazione con John e Yolande Helsop-Harrison e Ole Mattsson ai Royal Botanical Gardens, è riuscito a determinare parecchie caratteristiche biochimiche dello strato superficiale dello stimma delle angiosperme terrestri. Tra queste vi è la presenza di esterasi e un'affinità per la concanavalina *A*, una proteina trovata nel genere *Canavalia*, che appartiene al gruppo delle sostanze chiamate lectine. Si sa ora che la concanavalina *A* si combina con parecchi degli zuccheri trovati nei polisaccaridi e nelle glicoproteine, fra cui il mannosio e il glucosio. Recentemente, Adrienne Clarke e Paul Gleeson dell'Università di Melbourne hanno trovato che la lectina animale tridacnina, che è prodotta dal mollusco gigante *Tridacna maxima*, si lega anche con la superficie dello stimma di angiosperme terrestri. Si sa anche che la tridacnina ha affinità con determinati zuccheri.

La secrezione della superficie dello stimma di angiosperme terrestri comprende una componente zuccherina ed è chimicamente imparentata con le gomme vegetali. La sua funzione potrebbe essere quella di catturare e trattenere i granuli di polline dando così inizio al processo di germinazione.

Le secrezioni superficiali dello stimma di tutte le monocotiledoni marine che abbiamo esaminato assomigliano a quelle delle angiosperme terrestri per quanto riguarda la presenza di esterasi. Inoltre, la secrezione dello stimma di *Thalassia*

In questa fotografia al microscopio elettronico a trasmissione, lo stimma e il granulo di polline di *Amphibolis* sono ingranditi 15 000 volte. Il contatto tra la pellicola superficiale dello stimma, in basso, e quella del granulo di polline, in alto, dà luogo a una saldatura che unisce le strutture.

ha affinità per la concanavalina *A*. C'è tuttavia una differenza significativa: le secrezioni dello stimma delle piante terrestri sono solubili in acqua, mentre la secrezione delle monocotiledoni marine, una volta formata, non si disperde nell'acqua. La secrezione trattiene la sua integrità strutturale e, come hanno dimostrato gli esperimenti sulla impollinazione, quando la fecondazione è compatibile, la secrezione reagisce al contatto con il polline.

La ricerca sugli stimmi delle piante marine ha rivelato una correlazione morfologica fra lo stimma e le strutture del polline, che potrebbe essere facilmente riferita alla cattura di quest'ultimo, la quale dipende ovviamente dalla collisione casuale nell'ambiente marino. Gli stimmi, grazie alle piccole protuberanze che li rivestono, possono catturare sia i granuli di polline sferici coperti da sostanza viscosa, come avviene in *Thalassia* e in *Enhalus*, sia i filamenti di polline con tubi mucillaginosi come avviene in *Halophila*. Gli stimmi privi di prominenze, invece, catturano esclusivamente i granuli filiformi caratteristici di *Zostera marina* e del genere *Amphibolis*.

Fra le angiosperme terrestri la modificazione della superficie dello stimma influisce drasticamente sulla impollinazione. Per esempio, uno di noi (Knox), in collaborazione con Clarke all'Università di Melbourne, ha trovato che, quando si applica concanavalina *A* allo stimma di gladiolo, si impedisce l'impollinazione compatibile. Lo stimma perde la capacità di trattenere i granuli di polline di gladiolo e non avviene la penetrazione del tubetto pollinico nello stimma.

La scoperta suggerisce in primo luogo che, in condizioni normali, la secrezione superficiale dello stimma delle angiosperme terrestri è un efficace adesivo per il polline. Suggerisce inoltre che i siti superficiali dello stimma che legano la lectina svolgono un ruolo essenziale nel riconoscimento del granulo di polline. Ulteriori prove emergono da esperimenti compiuti dagli Heslop-Harrison e da Hugh Dickinson con *Agrostemma* e *Brassica* (il comune cavolo) nel laboratorio di Dickinson all'Università di Reading. Quando rimossero o modificarono la secrezione dello stimma con enzimi, il normale processo di impollinazione venne ostacolato.

Sorge così la domanda: la cattura, il trattenimento e la germinazione del polline avviene in maniera differente nel mondo marino? Lavorando insieme all'Università di Melbourne, preparammo in laboratorio colture di *Amphibolis* per studiare il processo di impollinazione, raccogliendo piante femminili e maschili dalle colonie del Bass Strait e trasferendole in acquari di acqua marina. I pollini liberati dalle piante maschili venivano raccolti e diretti sugli stimmi sommersi delle piante femminili. La maggior parte del polline non riusciva a incontrare uno stimma, ma tutti i granuli che venivano in contatto con esso immediatamente vi aderivano e venivano trattenuti.

A intervalli, i fiori impollinati venivano rimossi dalle piante femminili e gli stimmi venivano preparati per l'osservazione al microscopio. Furono eseguite sezioni sottili attraverso l'interfaccia polline-stimma che, all'esame con il microscopio elettronico, dimostrarono che il granulo di polline era strettamente attaccato allo stimma mediante un menisco di adesivo. Ciò era avvenuto probabilmente al momento del contatto grazie alla coalescenza dei due materiali superficiali; toccando la secrezione dello stimma, la pellicola di carboidrati e proteine posta sulla superficie del granulo di polline si combinava immediatamente con essa. Questo fenomeno sembra paragonabile a quello che avviene mescolando gli adesivi «a due componenti» in commercio.

Grazie a questa specificità di legame *Amphibolis* può discriminare fra il polline della propria specie e i pollini estranei. Se l'adesione infatti dipende dall'interazione di grandi molecole specifiche sulle superfici di contatto, il polline estraneo portatore di una pellicola superficiale con caratteristiche molecolari non adatte, semplicemente non può aderire allo stimma.

Un esperimento precedente aveva dimostrato che il polline di *Amphibolis* non riesce a germinare quando è incubato da solo in recipienti di acqua di mare. La scoperta è di particolare rilievo nell'interpretazione degli esperimenti, poiché si può concludere che il contatto con lo stimma è necessario per stimolare la germinazione del polline. In che forma avviene lo stimolo? Nelle angiosperme terrestri esso è veicolato dall'idratazione: quando il granulo di polline anidro si posa, assorbe acqua dallo stimma. Il granulo inumidito passa allora da uno stadio quiescente a uno attivato ed emerge il tubetto pollinico.

Nell'ambiente sottomarino è però assai improbabile che sia l'idratazione a dare il via alla germinazione. Per esempio, è dubbio che il grado di idratazione del granulo di polline sommerso possa anche solo avvicinarsi a quello riscontrato nel polline di terraferma. Naturalmente il contenuto di acqua del granulo di polline delle monocotiledoni marine può comunque essere ancora minore di quello dello stimma corrispondente.

Una possibile soluzione a queste difficoltà è stata fornita dal lavoro di A. W. A. M. de Cock dell'Università di Nimega. Studiando il polline di *Zostera marina*, de Cock trovò che un estratto del pistillo della pianta aggiunto a una soluzione zuccherina di acqua marina fa sviluppare meglio il tubetto pollinico. Egli suggerì che nell'estratto di pistillo vi fosse una molecola in grado di attivare lo sviluppo del tubetto pollinico. L'ipotesi ha interessanti implicazioni. Se una tale molecola promuove la crescita del polline della medesima specie, potrebbe anche giocare un ruolo nel riconoscimento della specie durante la fecondazione.

Se le molecole derivate dallo stimma richiedono il passaggio al granulo di polline per innescare la germinazione delle monocotiledoni marine, presumibilmente esse devono essere in grado di attraversare la parete del granulo. Come ciò avvenga dipende dalle loro dimensioni. Una sostanza di basso peso molecolare potrebbe essere trasmessa attraverso la parete per diffusione, mentre speciali condotti potrebbero facilitare il passaggio delle molecole più grandi. Abbiamo recentemente trovato tali condotti: attraversano la parete dei granuli di polline di *Heterozostera* e hanno un diametro di circa 100 nanometri.

Quando osservammo emergere il tubetto pollinico di *Amphibolis* trovammo veramente qualcosa di inaspettato. Nel polline delle angiosperme terrestri lo strato interno del granulo si estende per formare la prima parte del tubetto pollinico. Nel granulo di polline di *Amphibolis*, tuttavia, una regione della parete viene dis-

In queste fotografie eseguite al microscopio elettronico a trasmissione appaiono microstrutture della germinazione ingrandite rispettivamente 50 000 e 18 000 volte. La microfotografia a sinistra evidenzia alcuni piccoli canali presenti nella parete fibrillare del polline di *Heterozostera*. Possono servire come condotti per il passaggio dallo stimma al citoplasma del polline di sostanze stimolatrici della germinazione. Si vede anche la sottile pellicola che riveste la parete del polline. La microfotografia a destra mostra una sezione trasversale di un tubetto pollinico che sta crescendo verso il tessuto dello stilo di un pistillo dopo essere penetrato nella parete di una cellula recettiva dello stimma. Sopra il tubetto pollinico si nota la densa cuticola dello stimma attraversata dal tubo pollinico. Le strutture chiave sono riprodotte schematicamente in basso.

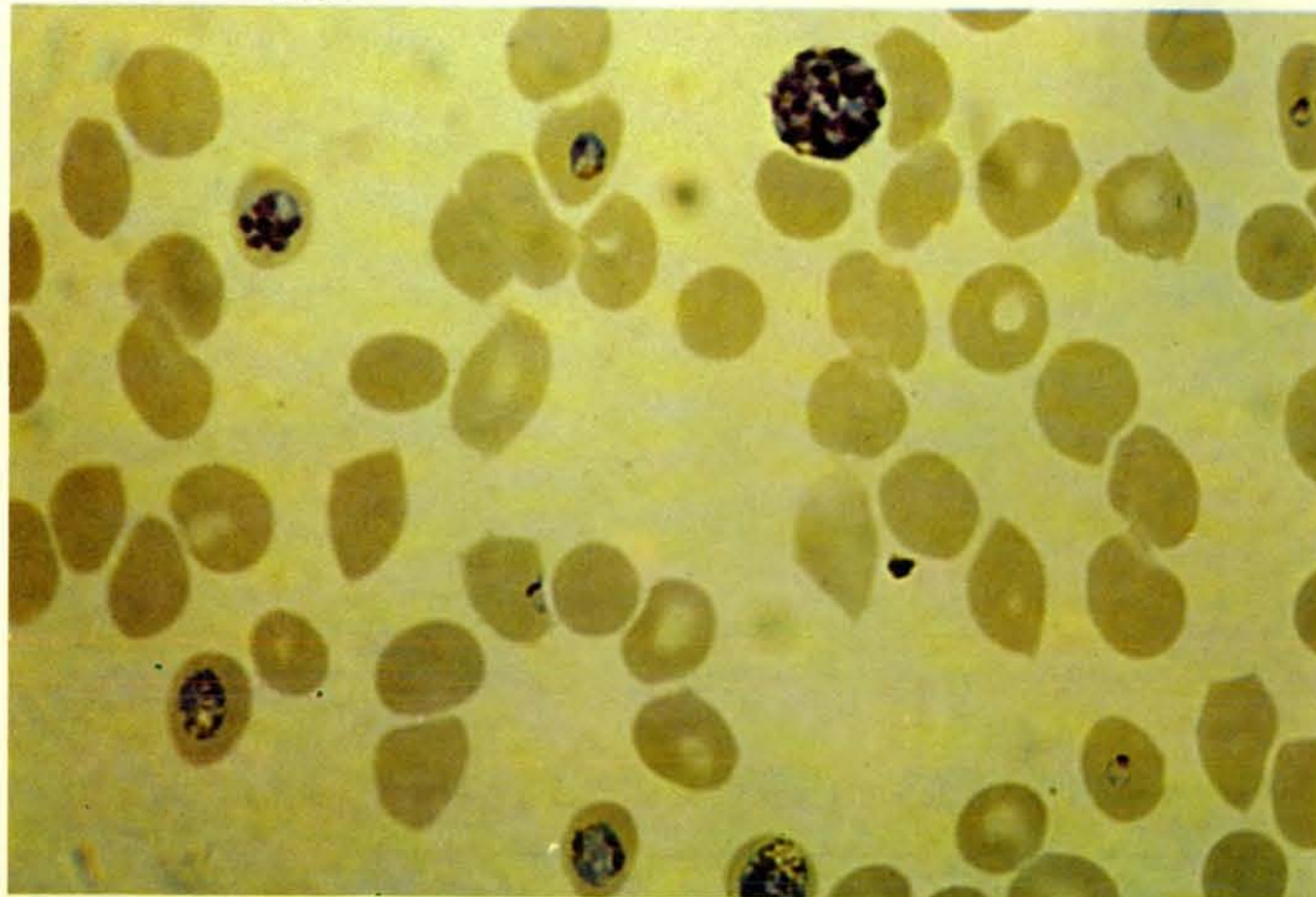
solta, apparentemente dall'azione enzimatica. Al microscopio l'apertura incipiente può essere rivelata come una piccola regione della parete del granulo che perde la sua affinità per certi coloranti. L'area è vicina al punto di attacco fra il granulo di polline e lo stimma, ma di fatto non si trova mai esattamente in quel punto. Questo indica che gli enzimi in grado di dissolvere la parete hanno origine nel granulo di polline piuttosto che nello stimma.

Nel polline di *Amphibolis* il passaggio di una bolla mucillaginosa attraverso l'apertura della parete del granulo di polline indica che la germinazione è incominciata. Sotto la bolla c'è una zona di citoplasma del polline densamente colorata, delimitata dalla membrana esterna del citoplasma. Questo passa attraverso l'apertura per costituire il vero e proprio tubetto pollinico. Non si sa ancora come si formino i materiali che danno origine alla parete del tubetto, ma sembra che i polisaccaridi della parete vengano sintetizzati frontalmente al citoplasma emergente.

Lo sviluppo del tubetto pollinico in *Amphibolis* differisce da quello delle angiosperme terrestri. In queste ultime il tubetto pollinico che si sviluppa è sempre strutturalmente continuo con lo strato interno di microfibrille che costituiscono la parete del granulo di polline. Una particolarità è stata recentemente trovata negli stadi iniziali della germinazione del polline di avocado (genere *Persea*). Margaret Sedgley dell'Australian Commonwealth Scientific and Industrial Research Organization di Adelaide ha osservato che i tubetti pollinici di questa pianta dapprima sono costituiti solo dalla membrana citoplasmatica esterna; la sintesi del materiale della parete del tubetto è ritardata fino a qualche tempo dopo che il tubetto ha preso contatto con lo stimma. Questa apparente somiglianza di sviluppo fra una pianta marina e l'avocado potrebbe dimostrarsi fortuita; tuttavia, l'esistenza dello stesso tipo di fenomeno in specie di piante così differenti incoraggia i tentativi di trovare una spiegazione funzionale.

Quando iniziarono le nostre osservazioni, l'interazione del granulo di polline con lo stimma non era stata osservata in alcuna pianta marina. Era chiaro che i nostri studi avrebbero potuto rivelare qualche estrema divergenza strutturale e fisiologica dagli schemi terrestri, ma era ragionevole aspettarsi che i principi generali dell'interazione polline-stimma potessero essere simili sia nell'ambiente marino che in quello terrestre. E così è stato.

Quali sono allora gli accorgimenti fisiologici che consentono l'impollinazione sottomarina? Per rispondere alla domanda occorre conoscere la precisa funzione della mucillagine del polline. Serve solo come adesivo? Regola il bilancio idrico del granulo di polline? O assolve entrambe le funzioni e inoltre, come la botanica inglese Agnes Arber suggerì più di mezzo secolo fa, impedisce alle sostanze solubili (di cui farebbero parte le molecole di riconoscimento) di diffondere nell'acqua?

Una coltura continua del parassita della malaria *Plasmodium falciparum* può essere fatta crescere in questo dispositivo. I parassiti vengono coltivati in uno strato sottile di globuli rossi umani che ricopre il fondo del tubo orizzontale ed è sommerso dal terreno nutritivo. Il contenuto di ossigeno del terreno e le altre condizioni sperimentali possono essere variate per mezzo dei tubi verticali. Le cellule vengono rimosse per l'analisi al microscopio attraverso il tubo corto al centro. Fino a che non furono stabilite le condizioni per la coltura del parassita, nel 1976, le ricerche sulla malaria provocata dal *P. falciparum*, la forma più letale, dovevano venire eseguite con il sangue di volontari umani o di primati.

Il destino dei parassiti della malaria viene controllato in strisci di globuli rossi su vetrini per microscopia. La colorazione (Giemsa) evidenzia i nuclei dei parassiti in viola scuro e il loro citoplasma in colore blu. La velocità di riproduzione viene misurata mediante conteggio dei numeri di parassiti per 100 globuli rossi. Questo striscio mostra i parassiti in vari stadi del loro ciclo vitale *(si veda l'illustrazione in basso a pagina 85).*

La biochimica della resistenza alla malaria

I geni di due malattie letali, l'anemia falciforme e la talassemia, sono favoriti dall'evoluzione perché proteggono contro la malaria. Il meccanismo di questa protezione ora può essere studiato in laboratorio

di Milton J. Friedman e William Trager

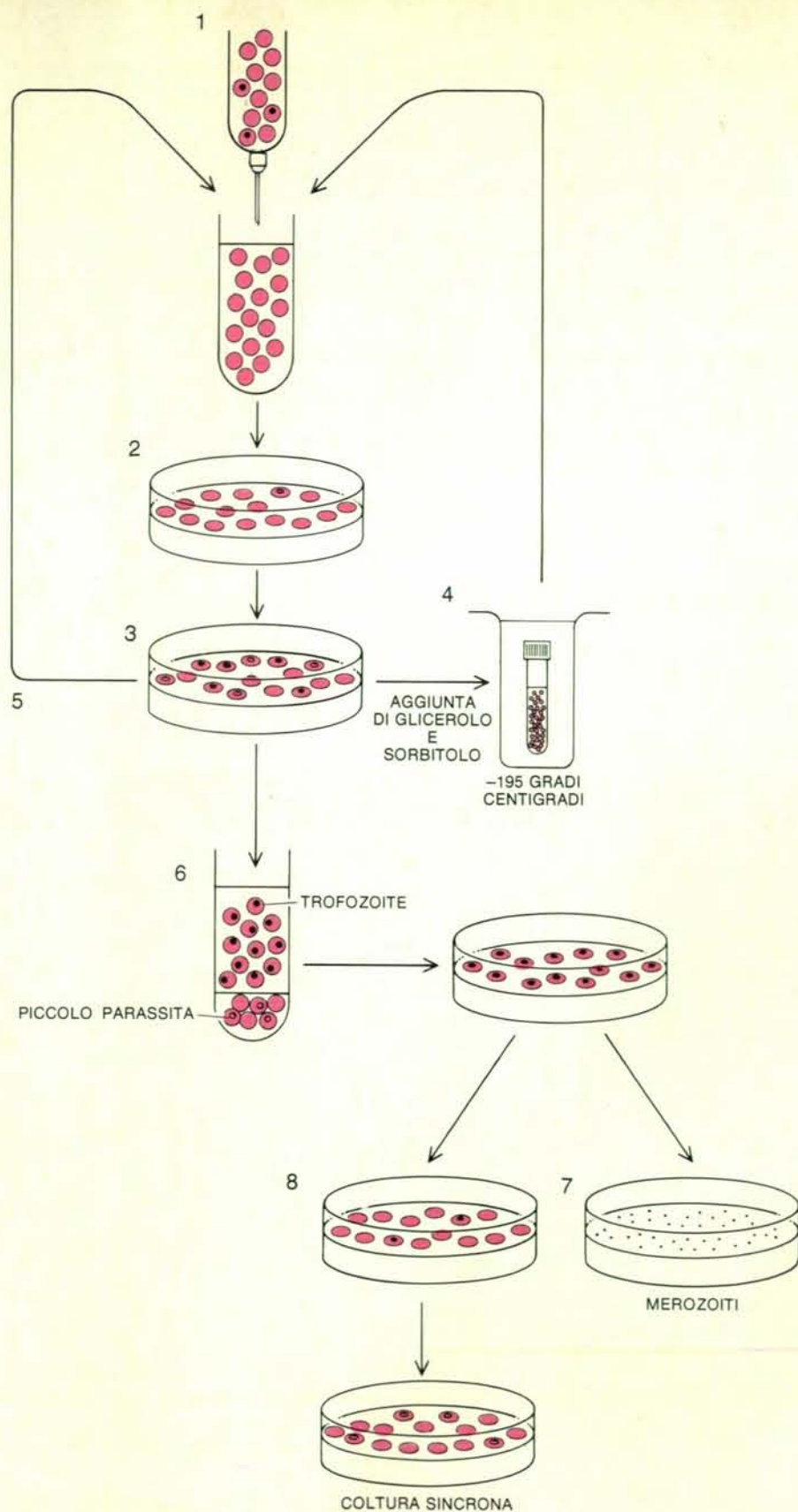
L'evoluzione è causata dalla selezione naturale, che opera sulla variabilità genetica prodotta dalla mutazione e dalla ricombinazione dei geni. I geni mutanti che conferiscono qualche vantaggio selettivo tendono a diventare più frequenti, mentre i mutanti svantaggiosi tendono a venire eliminati. Nelle popolazioni umane esistono pochi casi chiari di selezione a favore o contro geni specifici in risposta a forze di pressione selettiva. Gli esempi migliori sono costituiti dalle malattie ereditarie: la selezione agisce contro i geni che provocano queste malattie, e agisce tanto più efficacemente quanto più le malattie sono gravi.

Per questo motivo, le malattie ereditarie mortali dovrebbero essere molto rare. Tuttavia alcune anomalie ereditarie dei globuli rossi, in particolare la talassemia e l'anemia falciforme, in alcune popolazioni sono presenti con frequenze sorprendentemente alte. Questa è forse una prova contro la teoria della selezione naturale? Al contrario, i geni per l'anemia falciforme e la talassemia dimostrano l'efficacia della selezione come causa dell'evoluzione. Gli stessi geni mutanti che provocano l'anomalia mortale dei globuli rossi negli individui omozigoti (che posseggono due geni anormali, uno da ciascuno dei genitori) proteggono gli individui eterozigoti (che posseggono un gene anormale e uno normale) dagli effetti mortali della malaria, una malattia provocata da un parassita che infetta i globuli rossi. La protezione contro la malaria conserva le alte frequenze di questi geni per altro verso dannosi.

L'efficacia della malaria come forza selettiva deriva dal potente effetto di questa malattia parassitaria sulla salute e sulla capacità riproduttiva delle popolazioni umane. La malaria ha rappresentato storicamente una delle cause più frequenti di morte. In Africa la malaria è tuttora endemica: non colpisce la popolazione a ondate come le epidemie, ma piuttosto rappresenta una presenza costante che contribuisce a tassi di mortalità infantile che raggiungono il 50 per cento. Uccide circa il 10 per cento delle sue vittime direttamente e contribuisce alla morte delle altre diminuendo la capacità del sistema immunitario di combattere le altre infezioni. A causa della malaria un numero significativo di individui non vive abbastanza a lungo da riprodursi: qualsiasi mutazione genetica che renda resistenti alla malaria deve perciò avere un forte vantaggio selettivo.

Fu la coincidenza della distribuzione geografica dell'anemia falciforme con quella della malaria a suggerire la possibilità che il gene dell'anemia falciforme conferisse qualche tipo di resistenza. La dimostrazione clinica fu più difficile da ottenere, ma nel 1954 Anthony C. Allison dell'Università di Oxford dimostrò che i bambini eterozigoti per il gene dell'anemia falciforme venivano colpiti molto più raramente dalla forma più grave di malaria rispetto ai bambini che non possedevano questo gene. Tuttavia, poiché il parassita che induce la malaria non poteva venire mantenuto in colture di laboratorio, non fu possibile dimostrare la resistenza a livello cellulare, né fu possibile stabilire il suo meccanismo biochimico. Recentemente noi abbiamo sfruttato un sistema di coltura in vitro di nuova concezione che ci ha permesso di studiare in che modo il gene dell'anemia falciforme e alcuni altri geni mutanti che alterano la funzionalità dei globuli rossi conferiscono la resistenza alla malaria.

I globuli rossi, l'ambiente in cui il parassita della malaria viene in contatto con le alterazioni funzionali codificate da questi geni, sono costituiti in gran parte di emoglobina, la proteina che cattura l'ossigeno nei polmoni e lo cede ai tessuti. Le altre proteine del citoplasma dei globuli rossi sono enzimi metabolici. Alcuni di essi catalizzano la glicolisi, il processo in cui tramite passaggi successivi il glucosio viene degradato a lattato con formazione di due molecole di adenosintrifosfato (ATP). Altri catalizzano le reazioni di quella che è chiamata la via degli esosi monofosfati, che mantiene i coenzimi nicotinammide adenindinucleotide fosfato (NADP) e glutatione nella loro forma ridotta. L'ATP è il trasportatore universale di energia per la cellula; il NADP ridotto (NADPH) e il glutatione ridotto servono a prevenire e a riparare i danni provocati dall'ossidazione. La membrana cellulare racchiude la cellula e ne controlla la forma e la deformabilità. Inoltre controlla i movimenti degli ioni verso l'interno e l'esterno della cellula; in particolare mantiene, consumando ATP, un'alta concentrazione interna di potassio contro la tendenza verso un equilibrio con l'ambiente esterno povero di potassio. Sulla faccia esterna della membrana i glicolipidi e le glicoproteine caratterizzano in modo distintivo la superficie cellulare.

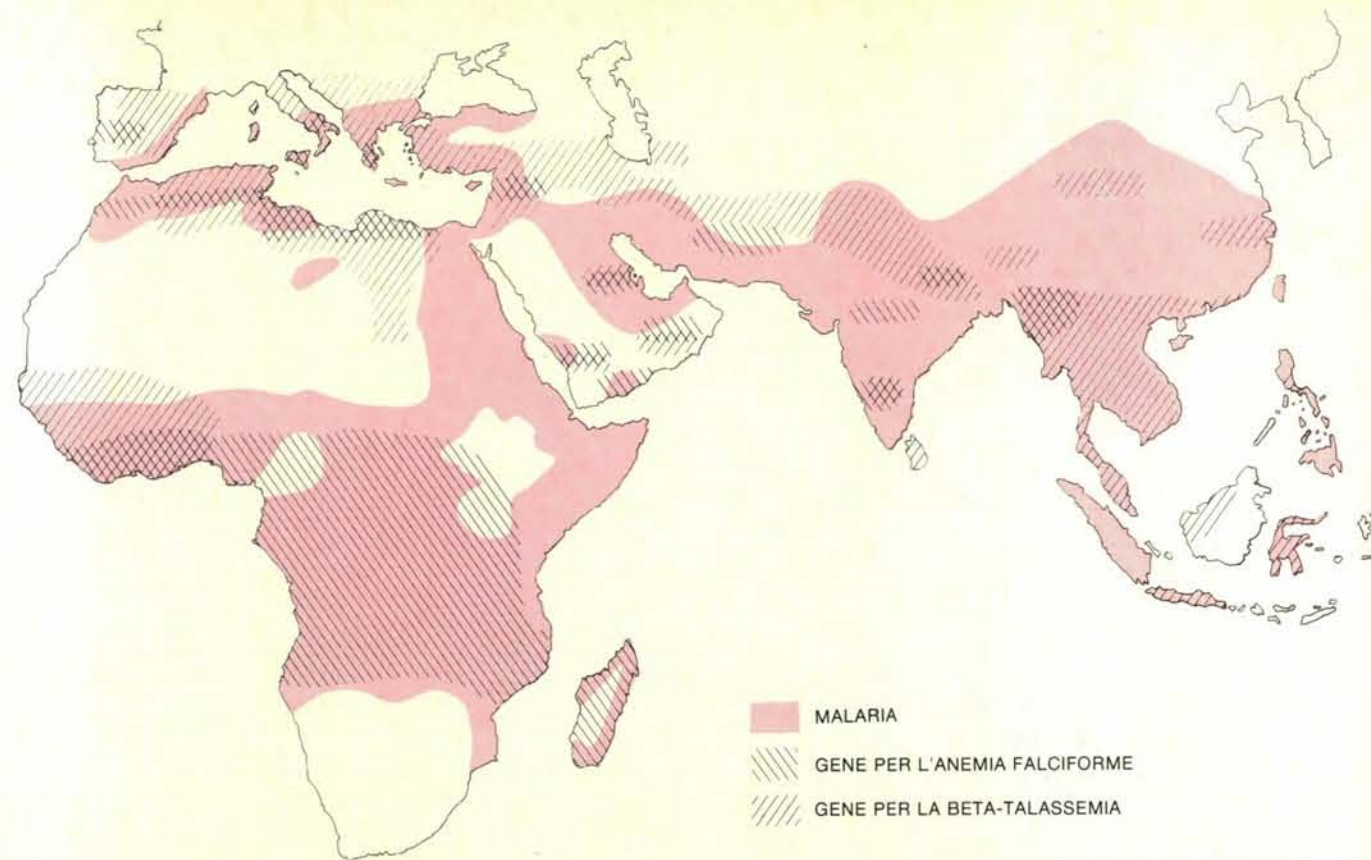
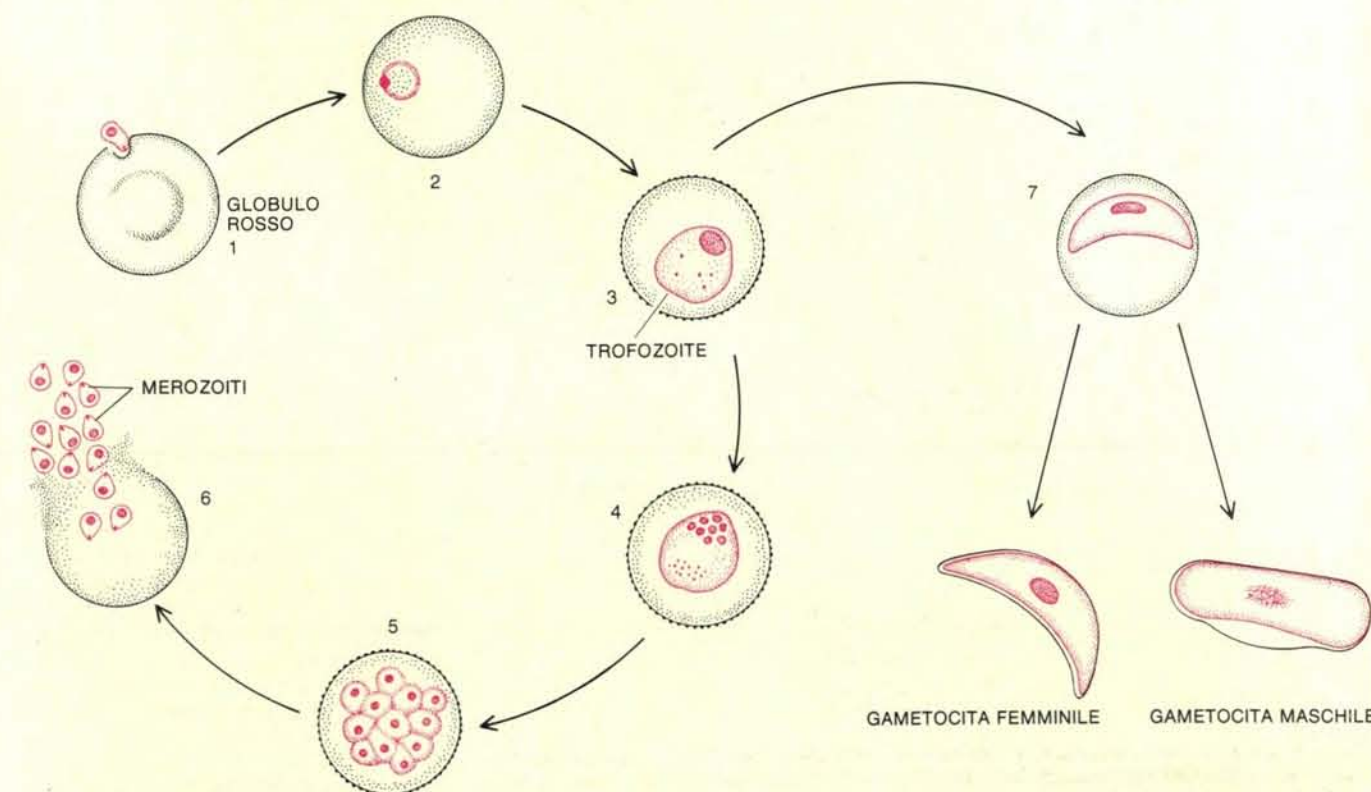
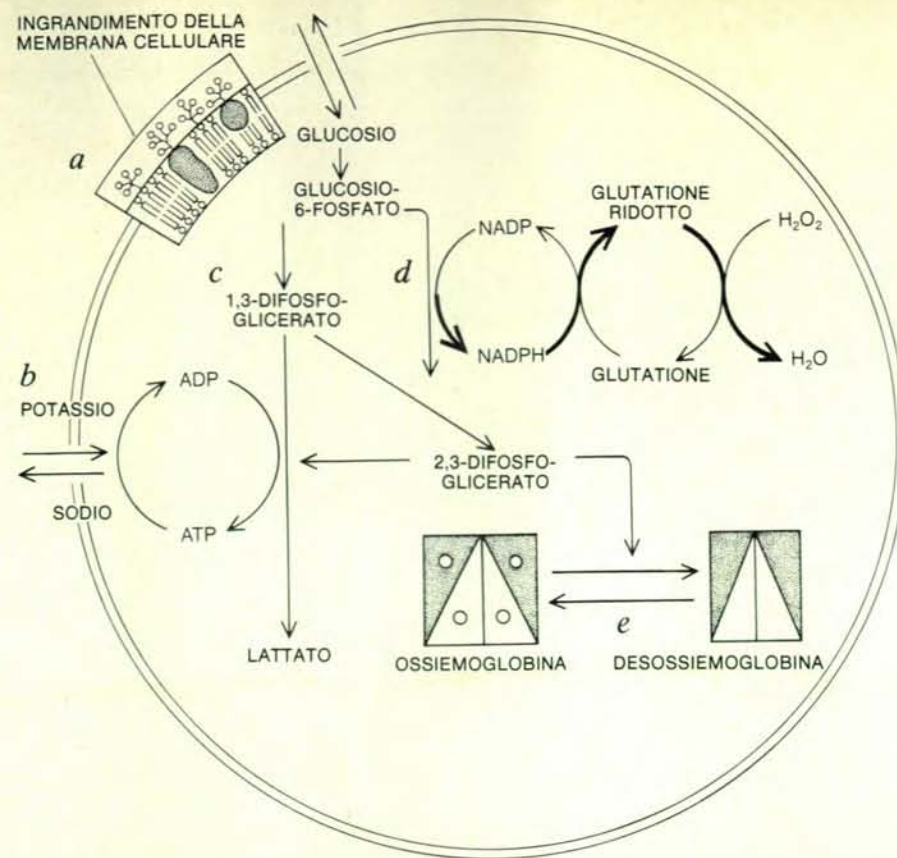
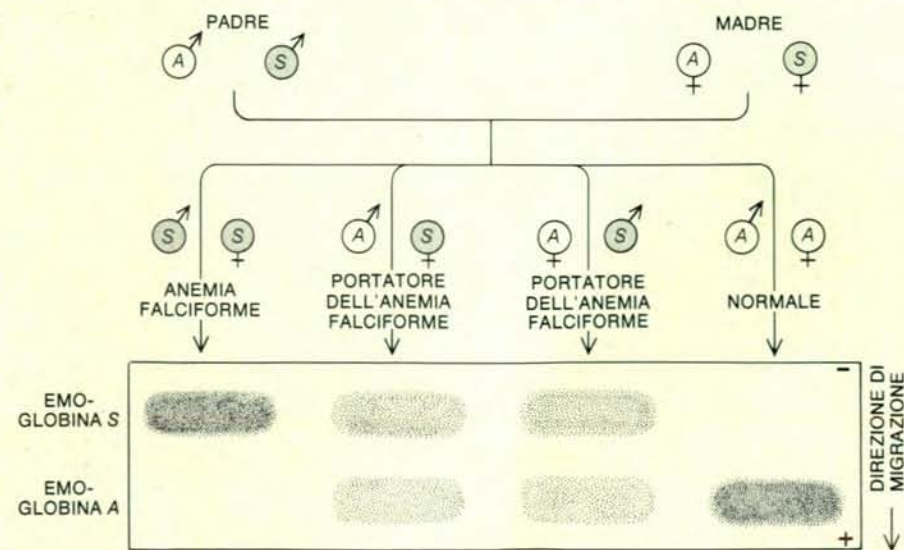
È a livello della membrana cellulare che il parassita della malaria interagisce inizialmente con la cellula. Il parassita è un piccolo protozoo unicellulare del genere *Plasmodium,* quattro specie del quale inducono la malaria nell'uomo; la forma più letale di malaria, che causa un milione di morti all'anno tra i bambini africani, è provocata dalla specie *Plasmodium falciparum.* Una forma specializzata del parassita viene iniettata nel sangue da una zanzara del genere *Anopheles* e migra nel fegato, dove si sviluppa e si divide a formare i merozoiti, la forma che infetta i globuli rossi. I merozoiti rientrano nel sistema circolatorio e riconoscono e si legano alle membrane dei globuli rossi. A questo punto si attiva un meccanismo, tuttora non ben conosciuto, per cui il merozoite attraversa la membrana cellulare, che gli si richiude intorno. Così racchiuso in un vacuolo il parassita cresce, digerendo l'emoglobina per ottenere gli amminoacidi che gli servono per produrre le proprie proteine, e presumibilmente utilizzando il glucosio, l'ATP e i coenzimi ridotti del globulo rosso per il proprio

Una coltura di *P. falciparum* si inizia inoculando una sospensione fresca di globuli rossi umani con globuli rossi infetti prelevati da un paziente o da una coltura di riferimento *(1)*. Le cellule vengono mantenute a temperatura corporea (su una piastra di terreno solido, come è mostrato in questo disegno, o nel tubo a U raffigurato a pagina 82) come uno strato sottile ricoperto di liquido nutritivo *(2)*. La coltura può venire saggiata per misurare la moltiplicazione dei parassiti. Dopo tre o quattro giorni le cellule contengono mescolanza di parassiti piccoli e grandi *(3)*, che possono venire congelati e conservati per usi futuri *(4)*, inoculati in nuove colture *(5)* o incubati in una soluzione di gelatina per separare i trofozoiti *(6)*; questi grossi parassiti possono venire coltivati da soli per produrre merozoiti *(7)* o mescolati con globuli rossi freschi per ottenere una coltura sincrona di parassiti *(8)*.

metabolismo. Dopo un periodo di crescita il nucleo del parassita si divide numerose volte; in seguito ogni nucleo viene circondato da membrane insieme a parte del citoplasma. In questo modo si formano da 12 a 24 nuovi merozoiti, che provocano la rottura del globulo rosso e invadono altre cellule.

La febbre periodica caratteristica della malaria è dovuta alla liberazione sincrona dei merozoiti e delle tossine prodotte dal parassita in tutto il corpo. Nel caso di *P. falciparum* questa liberazione sincrona si verifica ogni 48 ore, il periodo del ciclo di sviluppo dei parassiti nei globuli rossi. La febbre e la debilitazione che l'accompagna sono i sintomi principali della malaria. Nella malaria da *P. falciparum*, tuttavia, vi è un altro effetto più letale. Il globulo rosso infetto sviluppa delle sporgenze sulla sua superficie che ne provocano l'adesione alle pareti dei vasi capillari, dove rimane fino alla maturazione del parassita. Quando un gran numero di globuli rossi intasa in questo modo un organo vitale come il cervello, può aver luogo la morte del soggetto. L'indicazione più chiara dell'effetto protettivo del gene per l'anemia falciforme è che pochi portatori del gene muoiono a causa delle complicazioni cerebrali nella malaria da *P. falciparum*.

Linus Pauling e i suoi colleghi riuscirono a definire una malattia molecolare per la prima volta quando dimostrarono che nell'anemia falciforme la molecola di emoglobina è alterata, e che gli individui che ne sono affetti posseggono solo la forma alterata della molecola, chiamata emoglobina *S*, mentre alcuni membri delle loro famiglie posseggono circa metà emoglobina *S* e metà emoglobina normale *A*. Studi sulle famiglie, pubblicati circa nello stesso periodo, confermarono i risultati di Pauling, dimostrando che l'ereditarietà dell'anemia falciforme poteva essere ascritta a un solo gene recessivo: la malattia compare cioè solo nei membri della famiglia che sono omozigoti per quel gene.

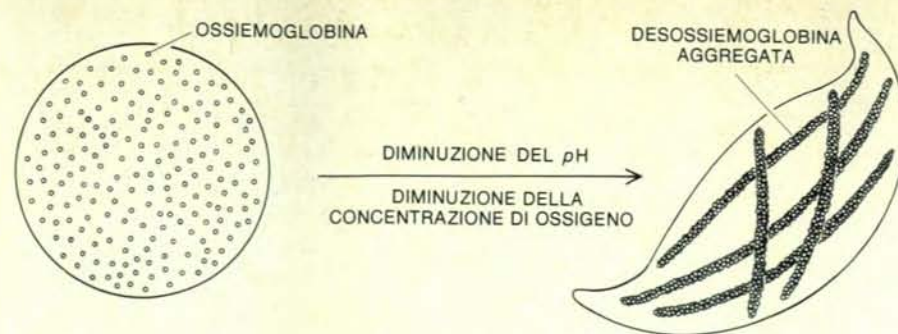
L'emoglobina adulta è costituita da due catene alfa e due catene beta; Vernon M. Ingram dell'Università di Cambridge dimostrò ben presto che nell'emoglobina *S* solo la catena beta è anormale. L'anomalia riguarda uno solo degli amminoacidi che costituiscono la catena: un acido glutammico sostituisce una valina. Gli individui omozigoti per il gene dell'anemia falciforme (designati *SS*) possiedono solo l'emoglobina *S* perché hanno entrambi i geni per la catena beta mutati; gli individui eterozigoti *(AS)*, che sono portatori del carattere per l'anemia falciforme, hanno un solo gene mutato, e circa il 40 per cento delle loro emoglobine sono di tipo *S*. Se due portatori del carattere per l'anemia falciforme hanno quattro figli, statisticamente uno su quattro sarà affetto da anemia falciforme, uno sarà completamente normale e due saranno portatori del carattere, in accordo con i principi della segregazione mendeliana.

I sintomi dell'anemia falciforme com-

Le distribuzioni dei geni per l'anemia falciforme e per la beta-talassemia coincidono con l'area dove la malaria provocata da *P. falciparum* era endemica prima del 1930 *(in colore)*. Questa coincidenza geografica ha fatto pensare che la resistenza alla malaria potesse essere il vantaggio evolutivo che tendeva a mantenere alta la frequenza dei geni responsabili di malattie letali del sangue in certe popolazioni umane.

Il ciclo vitale del *P. falciparum* nei globuli rossi inizia con l'invasione di una cellula da parte di un merozoite *(1)*. Il parassita incorpora una gocciolina di citoplasma, così che in sezione appare come un anello sottile *(2)*. La forma ad anello si nutre e cresce fino a diventare un trofozoite *(3)*; sulla superficie del globulo rosso si formano sporgenze che ne provocano l'adesione alle pareti dei vasi capillari. Lì il nucleo del parassita si divide ripetutamente *(4)*; ogni nucleo figlio acquista un poco di citoplasma *(5)*; il parassita si divide e dà origine a nuovi merozoiti, da 12 a 24, che provocano la rottura del globulo rosso ospite e danno inizio a un nuovo ciclo *(6)*. Alcuni trofozoiti si differenziano in gametociti maschili e femminili *(7)*; questi gametociti vengono ingeriti dalle zanzare e danno in tal modo inizio alla fase sessuale del ciclo vitale.

Un globulo rosso, la cellula più semplice dell'organismo, non è altro che citoplasma racchiuso da una membrana. I glicolipidi e le glicoproteine sulla sua superficie *(a)* determinano le interazioni della cellula con il suo ambiente e con il parassita della malaria. Le proteine della membrana controllano il trasporto di sostanze verso l'interno e l'esterno della cellula; in particolare, il potassio viene pompato verso l'interno e il sodio verso l'esterno *(b)*. Questa reazione richiede energia, che viene fornita dall'ATP generato durante la glicolisi, il processo con cui il glucosio viene degradato a lattato *(c)*. La via degli esosi monofosfati *(d)* produce NADP ridotto e glutatione ridotto, i quali prevengono e riparano il danno ossidativo a livello della membrana cellulare. Il citoplasma è riempito per la maggior parte con la proteina emoglobina *(e)*, che trasporta l'ossigeno dai polmoni ai tessuti per mezzo di un processo che si basa su cambiamenti strutturali delle quattro catene di amminoacidi che la compongono e che è controllato in parte dal difosfoglicerato.

Questo diagramma rappresenta la trasmissione dei geni per l'emoglobina anormale *S* da parte di genitori entrambi eterozigoti per il gene anormale. Statisticamente, uno dei figli sarà omozigote per il gene anormale *(SS)*, cioè eredita due geni per l'emoglobina *S* ed è affetto da anemia falciforme. Due figli erediteranno un gene per l'emoglobina *S* e un gene per l'emoglobina normale *A*; essi sono eterozigoti *(AS)* come i loro genitori e sono «portatori» del carattere per l'anemia falciforme. Un figlio sarà normale *(AA)*. Le emoglobine vengono identificate per mezzo dell'elettroforesi *(in basso)*. I campioni su un supporto di acetato di cellulosa sono sottoposti a una corrente elettrica. L'emoglobina *S* differisce dalla *A* per un solo amminoacido, che manca di una carica negativa, e così l'emoglobina *S* migra verso il polo positivo meno rapidamente della *A*.

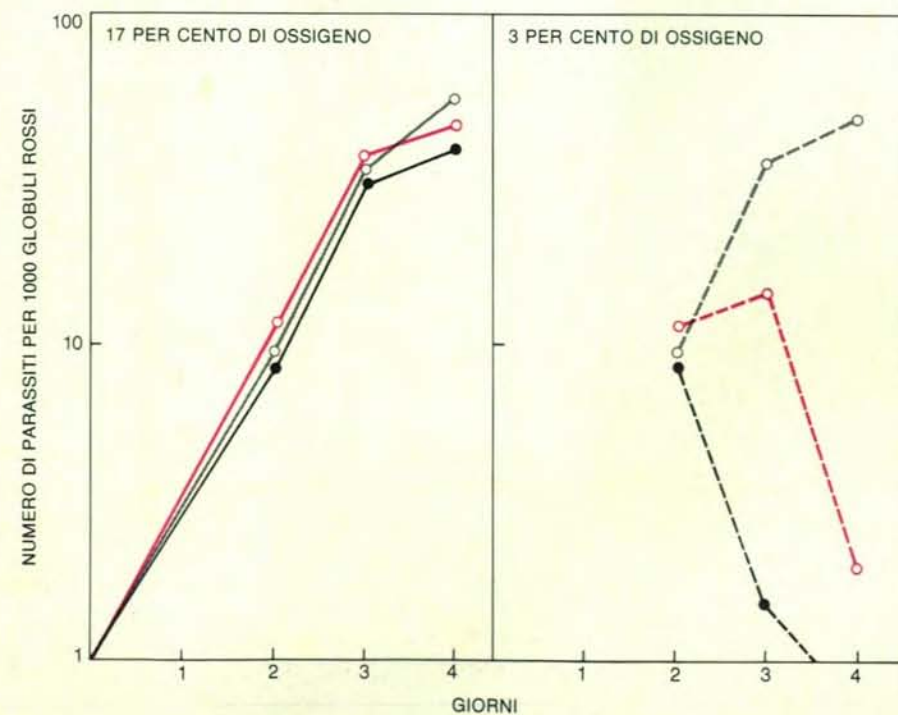
paiono quando i globuli rossi *SS* perdono ossigeno mentre circolano attraverso i tessuti. Quando l'emoglobina *S* è deossigenata, tende ad aggregarsi in lunghe fibre sottili, che modificano la forma normale a disco dei globuli rossi in forme angolose, compresa la caratteristica forma a falce che dà il nome alla malattia. In una crisi di anemia falciforme alcune cellule deformate bloccano la circolazione locale del sangue e impediscono l'afflusso di ossigeno. Quando la concentrazione di ossigeno diminuisce, altri globuli rossi si deformano e l'area di circolazione danneggiata si estende, causando intensi danni ai tessuti. In mancanza di cure mediche sofisticate la sopravvivenza degli omozigoti per l'emoglobina *S* è molto bassa. Gli individui portatori del carattere, invece, normalmente non presentano alcun disturbo. I loro globuli rossi *AS* contengono abbastanza emoglobina normale e si deformano solo in condizioni estreme, come a elevate altitudini.

Quattro anni fa, dopo mezzo secolo di tentativi di far crescere il parassita della malaria in colture di laboratorio, uno di noi (Trager) fu in grado di definire le condizioni che permettono il mantenimento di *P. falciparum* in colture continue di globuli rossi umani in un mezzo artificiale simile a sangue. Questo sistema di coltura viene attualmente utilizzato da alcuni ricercatori per sviluppare vaccini sperimentali contenenti materiale provenienti da vari stadi del ciclo cellulare del parassita, che vengono sperimentati sugli animali. Altri ricercatori stanno tentando di purificare e analizzare gli agenti biochimici che hanno effetti particolari nel corso di un'infezione. Il sistema ci ha anche permesso di studiare nei dettagli le interazioni del parassita della malaria con i globuli rossi mutanti ospiti.

Per capire come l'emoglobina dell'anemia falciforme protegge un portatore eterozigote contro la malaria abbiamo coltivato il parassita della malaria in globuli rossi prelevati da donatori normali, da individui portatori del carattere per l'anemia falciforme e da pazienti affetti da anemia falciforme. La coltivazione avveniva nelle nostre condizioni standard, in un'atmosfera con una concentrazione di ossigeno del 17 per cento, che crea nel mezzo di coltura una tensione di ossigeno simile a quella dei polmoni; l'emoglobina era perciò completamente ossigenata e i globuli rossi non si deformavano. In queste condizioni i parassiti crescevano ugualmente bene in tutti e tre i tipi di globuli rossi. Questo dimostra che non vi è alcuna alterazione sostanziale del metabolismo nei globuli rossi mutati e che la emoglobina *S*, come l'emoglobina *A*, può essere digerita dal plasmodio.

Per verificare l'effetto della deformazione sulla crescita del parassita aggiungemmo un piccolo numero di globuli rossi infettati a colture di globuli rossi *AS* e *SS* in una atmosfera di ossigeno al 17 per cento. Durante le 48 ore successive, il periodo di un ciclo di crescita, tutti i parassiti abbandonarono i globuli rossi normali in cui

L'emoglobina *S* nel suo stato ossigenato *(a sinistra)* è dispersa all'interno del globulo rosso, che ha una normale forma a disco. A differenza dell'emoglobina *A*, tuttavia, l'emoglobina *S* tende ad aggregarsi quando cede ossigeno ai tessuti, formando strutture aghiformi quasi cristalline che deformano il globulo rosso in una forma rigida e angolosa *(a destra)*. Le cellule deformate possono intasare i capillari e diminuire il flusso sanguigno, in questo modo producono una diminuzione della concentrazione di ossigeno e promuovono quindi la deformazione di un maggior numero di cellule.

erano cresciuti e invasero i globuli rossi mutanti. Quando abbassammo la concentrazione di ossigeno al 3 per cento i globuli rossi *SS* subirono la caratteristica deformazione, così come alcuni dei globuli rossi *AS*. Seguimmo la crescita dei parassiti giorno per giorno. Dopo un giorno in condizioni di bassa concentrazione di ossigeno quasi nessun parassita era visibile nelle colture di globuli rossi *SS:* si erano lisati, cioè disintegrati, insieme ai globuli rossi che li ospitavano. Nelle colture di globuli rossi *AS*, d'altra parte, il numero di parassiti vivi diminuì significativamente solo al secondo giorno, e invece che disintegrarsi i parassiti uccisi rimanevano visibili come masse raggrinzite all'interno dei globuli rossi. Il loro aspetto ci ricordava quello che presentano i plasmodi morti per mancanza di nutrimento, forse a causa di qualche inibizione metabolica provocata indirettamente dalla deformazione dei globuli rossi.

Se questo era vero, i parassiti avrebbero dovuto essere protetti impedendo la deformazione dei globuli rossi. Trattammo i globuli rossi *AS* con cianato, che aumenta l'affinità dell'emoglobina *S* per l'ossigeno, rendendo più improbabile l'aggregazione a basse concentrazioni di ossigeno.

La moltiplicazione di *P. falciparum* è circa la stessa in cellule normali *(cerchi vuoti)*, in cellule *AS* *(cerchi in colore)* e in cellule *SS* *(cerchi neri pieni)* finché le cellule si trovano a una concentrazione di ossigeno del 17 per cento *(a sinistra)*. Se dopo due giorni tale concentrazione viene abbassata al 3 per cento, tuttavia, solo i parassiti nelle cellule normali continuano a moltiplicarsi *(a destra)*. Quelli che infettano le cellule *SS* muoiono in un giorno, quelli che infettano le cellule *AS* in due.

Dopo il trattamento con cianato e un lavaggio, i globuli rossi *AS* erano ancora in grado di ospitare il *Plasmodium falciparum*, ma non si deformavano più così facilmente. Quando i globuli rossi così trattati furono infettati e coltivati a una concentrazione di ossigeno del 3 per cento, i parassiti sopravvissero. L'inibizione che era stata osservata nei globuli rossi non trattati doveva perciò essere dovuta alla deformazione delle cellule, non semplicemente alla bassa concentrazione di ossigeno. In che modo la deformazione dei globuli rossi poteva inibire la crescita dei parassiti?

Una delle conseguenze della deformazione dei globuli rossi contenenti emoglobina *S* è che la membrana cellulare diventa più permeabile al potassio, che in questo modo viene perduto dalla cellula; nelle condizioni di bassa concentrazione di ossigeno il contenuto di potassio delle nostre cellule ospiti *AS* diminuiva. Era stato precedentemente dimostrato che i parassiti conservati all'esterno dei globuli rossi richiedono un ambiente ad alta concentrazione di potassio, e così ipotizzammo che la perdita di potassio dovuta alla deformazione avrebbe potuto essere la causa dell'inibizione del metabolismo dei parassiti. Mettemmo alla prova la nostra ipotesi coltivando i globuli rossi *AS* infettati in un terreno colturale con solo il 3 per cento di ossigeno, ma con un'alta concentrazione di potassio. Le cellule subirono la solita deformazione dovuta alla mancanza di ossigeno, ma questa volta il contenuto di potassio rimase elevato e i parassiti sopravvissero. In altre parole, era possibile impedire l'inibizione della crescita del parassita impedendo la perdita di potassio da parte delle cellule *AS* deformate. (In condizioni uguali i plasmodi coltivati nelle cellule *SS* non erano protetti e morivano per lisi. Le fotografie al microscopio elettronico ci indicarono il perché: dopo sei ore di deossigenazione era possibile osservare fasci appuntiti di emoglobina *S* aggregata che trafiggevano alcuni plasmodi; le membrane di altri parassiti erano state distrutte e i plasmodi erano parzialmente lisati. In altre parole i plasmodi nelle cellule *SS* non morivano a causa dell'inibizione metabolica, ma per reale rottura fisica.)

La sequenza degli eventi nelle cellule *AS* sembrava quindi essere la seguente: la deformazione abbassava la concentrazione di potassio intracellulare, e questa a sua volta uccideva i parassiti. Questi risultati tuttavia non erano sufficienti a spiegare completamente la resistenza dell'eterozigote alla malaria. Poiché la cellula *AS* contiene meno emoglobina *S* della cellula *SS*, non si deforma in condizioni naturali; la sua circolazione attraverso le regioni del corpo a bassa concentrazione di ossigeno è troppo veloce per permettere la sua deformazione. La circolazione di una cellula infettata è impedita in larga misura dalle piccole sporgenze della membrana cellulare provocate dai plasmodi, e questo trattiene la cellula in regioni a bassa concentrazione di ossigeno per parecchie

ore; anche così, tuttavia, meno del 5 per cento delle cellule infettate subirebbe la caratteristica deformazione, se non intervenisse un altro effetto dell'infezione.

Lucio Luzzatto e i suoi colleghi dell'Istituto internazionale di genetica e biofisica di Napoli hanno dimostrato che le cellule infettate si deformano molto più facilmente delle altre. Perché? Noi abbiamo trovato che l'ambiente intracellulare di una cellula infettata è più acido (il *p*H è 0,4 unità più basso) che in una cellula non infettata, e questo aumenta significativamente la velocità di deformazione. Abbiamo calcolato, inoltre, che il *p*H più basso delle cellule infettate aumenta anche la frequenza delle deformazioni, fino al 40 per cento in più.

Nel loro complesso queste osservazioni suggeriscono il seguente meccanismo di protezione contro la malaria negli individui eterozigoti per l'anemia falciforme. Il parassita all'interno della cellula *AS* si sviluppa normalmente finché la cellula non viene bloccata all'interno dai tessuti. Qui, a causa della bassa concentrazione di ossigeno e del *p*H intracellulare basso, la cellula ospite si deforma, la concentrazione di potassio cade e il parassita muore. Questa sequenza di eventi può proteggere efficacemente dalla malaria anche se non tutti i parassiti vengono uccisi in questo modo, perché una riduzione della velocità di riproduzione del plasmodio assicura al sistema immunitario il tempo necessario per sviluppare una risposta immunitaria protettiva. (È anche possibile immaginare un'ipotesi alternativa. Le cellule infettate potrebbero, per qualche ragione a noi ignota, deformarsi mentre sono in circolo invece che solo quando sono bloccate nei tessuti, e potrebbero venire eliminate dall'azione filtrante della milza. La prima ipotesi tuttavia è confermata dal fatto che gli eterozigoti non sono protetti contro quei tipi di malaria in cui le cellule infettate non formano sporgenze sulla loro superficie e non vengono bloccate nei tessuti.)

Oltre all'anemia falciforme ci sono altre anomalie genetiche dei globuli rossi la cui distribuzione geografica è stata correlata a quella dell'incidenza della malaria, suggerendo che i geni responsabili di queste alterazioni possono anch'essi conferire qualche tipo di resistenza. Tra queste malattie genetiche vi sono le talassemie, che sono causate da una deficienza nella produzione dell'una o dell'altra catena di emoglobina. La beta-talassemia, per esempio, è provocata da una insufficiente sintesi di catena beta. Gli omozigoti beta-talassemici sono affetti dalla cosiddetta anemia di Cooley, una grave condizione patologica per cui non viene sintetizzata, se non in misura minima, la forma adulta normale dell'emoglobina; la vita dei pazienti può essere prolungata unicamente per mezzo di continue trasfusioni di sangue. Malgrado ciò, in molte regioni malariche, e in particolare sulle sponde del Mediterraneo, circa l'1 per cento dei bambini sono omozigoti per il gene della beta-talassemia e sono affetti da anemia di Cooley; gli eterozigoti sono sani. La resistenza alla malaria non è stata convincentemente dimostrata per gli eterozigoti beta-talassemici, ma la correlazione geografica tra la frequenza del gene e l'incidenza storica della malaria è impressionante. Per citare un solo esempio, la frequenza genica è molto più elevata nelle valli della Sardegna, dove la malaria è stata per lungo tempo endemica, rispetto alle zone di montagna, dove l'incidenza della malaria era bassa.

In queste microfotografie si vedono globuli rossi infettati dai parassiti coltivati per un giorno in presenza di un'atmosfera di ossigeno al 3 per cento e fissati in formaldeide per conservarne la forma *(in alto)* o strisciati su un vetrino per una migliore osservazione microscopica *(in basso)*. Le cellule normali *AA (a sinistra)* sono a forma di disco; alcune delle cellule *AS (al centro)* e tutte le *SS (a destra)* si sono deformate. I parassiti nelle cellule *AA* sono vivi, quelli nelle cellule *AS* sono morti *(freccia)* o moriranno entro le 24 ore. Nelle cellule *SS* non si vedono parassiti.

Ci proponemmo di dimostrare la resistenza del nostro sistema cellulare *in vitro*. Sapevamo che una caratteristica delle cellule talassemiche è un'anormale sensibilità della membrana cellulare al danno ossidativo. Quando una molecola viene ossidata, perde alcuni elettroni che altrimenti sarebbero disponibili per formare legami chimici. Quando le molecole lipidiche (grassi), che sono le principali componenti della membrana, sono ossidate, si frammentano e compromettono l'integrità della membrana. Gli agenti che causano l'ossidazione nelle cellule non sono stati tutti completamente identificati, ma è noto che uno di questi agenti è rappresentato dal perossido di idrogeno («acqua ossigenata»).

Il parassita della malaria produce perossido di idrogeno all'interno della cellula ospite, come è stato dimostrato da N. Etkin e John W. Eaton della University of Minnesota Medical School. I perossidi causano in tutte le cellule danni ossidativi e costituiscono un serio problema per il mantenimento dell'integrità della cellula. Nei globuli rossi talassemici ipersensibili il perossido di idrogeno potrebbe causare danni alla membrana. Confermammo questa ipotesi dimostrando che i parassiti in eterozigoti all'interno di cellule beta-talassemiche sono più sensibili dei parassiti ospiti di cellule normali a tre condizioni sperimentali. La prima condizione è un ambiente ad alta concentrazione di ossigeno, dal 25 al 30 per cento. La seconda è la presenza di certe sostanze chimiche che catalizzano le reazioni di ossidazione. L'ultima è l'assenza dal terreno colturale di un componente normale, il glutatione ridotto, che è un intermedio della via metabolica che riduce la concentrazione di perossido di idrogeno all'interno della cellula.

TERRENO NORMALE, 17 PER CENTO DI OSSIGENO — POTASSIO — SODIO
TERRENO NORMALE, 3 PER CENTO DI OSSIGENO — *AA* — *AS* — *SS*
ELEVATA CONCENTRAZIONE DI POTASSIO, 17 PER CENTO DI OSSIGENO
ELEVATA CONCENTRAZIONE DI POTASSIO, 3 PER CENTO DI OSSIGENO — *AS* — *SS* — *AA*

La perdita di potassio è la causa della morte del parassita nelle cellule *AS*. Alla concentrazione di ossigeno del 17 per cento *(a sinistra)* la membrana cellulare rimane intatta e la concentrazione di potassio all'interno della cellula viene mantenuta a livello fisiologico sia in un terreno colturale povero di potassio *(in alto)* sia in uno ad alta concentrazione di potassio *(in basso)*. In condizioni di bassa concentrazione di ossigeno e scarsità di potassio *(in alto a destra)* le cellule *AS* e *SS* si deformano; le loro membrane vengono danneggiate e non trattengono più il potassio; i parassiti si accartocciano *(AS)* o scompaiono *(SS)*. L'incubazione in un terreno colturale caratterizzato da una elevata concentrazione di potassio *(in basso a destra)* protegge i parassiti che infettano le cellule *AS* malgrado la deformazione e la distruzione delle loro membrane cellulari, ma non protegge i parassiti che si trovano all'interno delle cellule *SS*. La causa della morte dei parassiti nelle cellule *AS*, allora, è la mancanza di un fattore metabolico essenziale, il potassio. Nelle cellule *SS*, d'altra parte, gli aggregati di emoglobina *S* distruggono i parassiti trafiggendoli meccanicamente.

Abbiamo calcolato che ognuna di queste tre condizioni aumenta il carico di danni ossidativi alla cellula. Eaton ha trovato che nei topi infettati con la malaria il carico ossidativo e la protezione contro di esso sono finemente bilanciati. Ogni ulteriore sensibilità in una cellula talassemica può quindi influire sull'esito dell'infezione. (Sebbene una cellula talassemica contenga meno emoglobina, questa è apparentemente sufficiente a fornire il nutrimento al parassita; la bassa concentrazione di emoglobina di per sé non inibiva la moltiplicazione dei parassiti nel nostro sistema sperimentale.)

Abbiamo trovato che la vitamina E, che protegge i lipidi della membrana cellulare contro l'ossidazione, evita la morte dei parassiti contenuti in cellule di un portatore eterozigote talassemico in tutte le condizioni sperimentali. Questo è in accordo con l'idea che la membrana di cellula eterozigote beta-talassemica venga danneggiata dall'ossidazione nel corso dell'infezione malarica. Per quanto riguarda il meccanismo, di nuovo sembra che sia implicato il potassio. In un terreno colturale contenente un'alta concentrazione di potassio i parassiti si sviluppano normalmente in cellule eterozigoti beta-talassemiche in tutte le condizioni che invece ne causano la morte in un terreno colturale con una bassa concentrazione di potassio.

Le nostre ricerche hanno prodotto tre interessanti risultati aggiuntivi. Uno è stato la conferma di una correlazione a lungo sospettata tra la malaria e una deficienza ereditaria del contenuto dell'enzima glucosio-6-fosfato deidrogenasi nei globuli rossi (G6PD). Anche questo carattere genetico si trova prevalentemente nelle zone malariche; studi sul campo hanno in qualche caso, ma non sempre, dimostrato una correlazione con la resistenza alla malaria. Il G6PD è il primo enzima della via degli esosi monofosfati, che rigenera l'NADPH, un coenzima essenziale per la protezione contro il danno ossidativo. Sembrava che i globuli rossi mancanti di G6PD, come quelli eterozigoti per la beta-talassemia, fossero più sensibili al perossido di idrogeno generato dai parassiti della malaria. Noi abbiamo trovato che in effetti i parassiti che infettano le cellule carenti di G6PD sono molto sensibili al danno ossidativo e sono protetti dagli agenti antiossidanti.

Un altro dei risultati delle nostre ricerche sui globuli rossi talassemici riguarda il favismo, un'anemia emolitica promossa dall'ingestione di fave, un alimento consumato in tutto il bacino del Mediterraneo. Le fave contengono sostanze che potrebbero aumentare la sensibilità dei globuli rossi all'ossidazione; alcune di queste sostanze sono correlate ai catalizzatori ossidativi che noi avevamo utilizzato nei nostri esperimenti. Se i nostri esperimenti simulavano l'ingestione di fave da parte degli individui beta-talassemici o carenti di G6PD nei globuli rossi, questo voleva dire che il consumo di fave (e forse anche di altri cibi non ancora identificati) aumenta il livello di protezione contro la malaria negli individui eterozigoti per queste due anomalie genetiche dei globuli rossi. Questo effetto della dieta spieghe-

rebbe anche i risultati inconcludenti degli studi sulla resistenza alla malaria di questi eterozigoti.

Il terzo risultato accessorio riguarda i neonati. Durante i primi mesi di vita i neonati sono quasi completamente protetti contro la malaria. Geoffrey Pasvol, R. J. M. Wilson e D. J. Weatherall dell'Università di Oxford hanno recentemente dimostrato che l'emoglobina fetale (che consiste di catene alfa e gamma, persiste per un certo tempo dopo la nascita e si trova anche nei globuli rossi di alcuni adulti) può contribuire a questa protezione; anche i globuli rossi di individui adulti erano resistenti all'infezione da parte dei plasmodi se contenevano emoglobina fetale. Abbiamo dimostrato che i parassiti che si trovano all'interno dei globuli rossi che contengono emoglobina fetale sono molto sensibili al danno ossidativo, come quelli che infettano i globuli rossi talassemici. In ogni caso è improbabile che la malaria sia stata una forza selettiva nell'evoluzione dell'emoglobina fetale. La proteina fetale ha una più alta affinità per l'ossigeno dell'emoglobina adulta, e ciò migliora il rifornimento di ossigeno al feto in sviluppo: questa è stata probabilmente la forza selettiva che ne ha determinato l'evoluzione. La sensibilità agli ossidanti dei globuli rossi fetali e la conseguente resistenza alla malaria sono probabilmente effetti secondari di questo adattamento alle condizioni dello sviluppo del feto.

La sequenza di eventi qui raffigurata potrebbe proteggere gli eterozigoti *AS* contro la malaria. Una cellula infettata dal parassita è caratterizzata dalla presenza di sporgenze sulla sua superficie e da un *p*H intracellulare basso. Una cellula non infettata *AS* trascorrerà un breve periodo di deossigenazione in un capillare senza deformarsi *(1)*; una cellula infettata, d'altra parte, rimarrà bloccata abbastanza a lungo da permettere alla bassa ossigenazione e al *p*H basso di provocare la deformazione *(2)*. La deformazione causa la perdita di potassio attraverso la membrana cellulare *(3)*. Privato di potassio, il parassita muore *(4)*. La morte di una frazione dei parassiti può concedere all'organismo dell'eterozigote il tempo sufficiente per lo sviluppo di una risposta immunitaria.

Per riassumere, l'evoluzione ha prodotto la selezione di globuli rossi mutanti che funzionano sufficientemente bene in condizioni normali, ma si danneggiano quando sono infettati dal *Plasmodium falciparum*, e questo danno uccide il parassita. In altre parole, questi globuli rossi sono marginalmente vitali e l'infezione li rende non vitali e incapaci di ospitare il parassita intracellulare. L'alterazione genetica nel caso delle cellule *AS* è una tendenza appena contenuta a deformarsi, che è stimolata dalla presenza del parassita. L'alterazione dei globuli rossi eterozigoti talassemici consiste in un aumento della sensibilità dei lipidi della membrana cellulare nei confronti del perossido di idrogeno generato dal parassita. In entrambi i casi l'effeto finale è la perdita di potassio che inibisce il metabolismo del parassita.

La storia evolutiva di un gene può essere molto complessa. Quando una nuova mutazione casuale si manifesta, non è possibile prevedere il suo effetto o il suo potenziale evolutivo. Sicuramente non si sarebbe potuto prevedere che certi geni avrebbero causato malattie letali negli omozigoti e protetto gli eterozigoti dalla malaria. Il contrasto tra le due opposte forze selettive della malaria da una parte e dell'anemia falciforme e della beta-talassemia dall'altra parte ha prodotto un polimorfismo bilanciato: una situazione in cui il vantaggio dell'eterozigote combinato con lo svantaggio dell'omozigote mantiene il gene mutante a un livello di frequenza nella popolazione basso ma consistente.

I meccanismi della vita e della morte si fondano sulla chimica e sulla biologia molecolare; possono essere riprodotti e compresi, almeno in parte, in laboratorio. I loro effetti finali sulla specie umana non si decidono, tuttavia, né in laboratorio né nelle singole cellule degli esseri umani ma, lentamente e imprevedibilmente, nella lotta evolutiva per la vita combattuta sulla scala dei continenti e dei millenni.

Particolare di un pozzo esplorativo per la prospezione di gas naturale, installato a Settala nella Pianura Padana. Questa indagine, tendente ad appurare la consistenza di giacimenti di gas naturale, è stata terminata nel 1980, ma non ha fornito valori incoraggianti data la scarsa concentrazione di gas riscontrata. Non si sa ancora se questo, come altri pozzi disseminati in Lombardia, potrà essere economicamente sfruttabile.

Contaminazione ambientale da fonti energetiche

Non è possibile stabilire a priori il grado di contaminazione ambientale derivante dalle diverse attività umane: il pensare, comunque, ad attività industriali che operino con impatto ambientale uguale a zero è un'utopia

di Mario Dall'Aglio

L'uomo è senza dubbio l'animale che interagisce in maniera più marcata con l'ambiente nel quale vive. Sono millenni che la natura viene sistematicamente alterata da questa specie animale molto attiva che, dopo aver eliminato la concorrenza degli altri animali per il dominio del territorio, ha fatto aumentare esponenzialmente le proprie esigenze, e quindi il flusso di materie prime e manufatti a essa «necessari».

Solo in tempi recenti le profonde modificazioni indotte sull'ambiente dalle attività umane in genere, e industriali in particolare, hanno ricevuto l'attenzione che meritano da parte della comunità nazionale, delle forze politiche e sociali e delle comunità locali. In tempi recentissimi l'attenzione generale è stata indirizzata in particolar modo all'impatto ambientale delle fonti energetiche; soprattutto l'energia nucleare, in tutte le fasi del ciclo produttivo - dalla estrazione mineraria dell'uranio alla sistemazione ultima dei residui radioattivi - è stata messa sotto processo da parte di comunità locali, e di alcune forze politiche, per i potenziali danni all'ambiente e all'uomo che essa sarebbe in grado di produrre nel breve come nel lungo periodo. L'argomento ha appassionato l'opinione pubblica in tutto il mondo occidentale, e già in numerose occasioni la decisione relativa alla messa in funzione di centrali elettronucleari è passata attraverso il vaglio di referendum popolari (California, Austria, Svizzera, Svezia).

Ma un vivo interesse è stato dedicato anche ad altre fonti energetiche. Non va infatti dimenticato che in Italia la prima massiccia contestazione popolare si è avuto molti anni fa (nel 1971) contro l'installazione a Piombino di una centrale a olio combustibile. Le attuali profonde preoccupazioni delle popolazioni dei Campi Flegrei per l'esplorazione geotermica condotta dall'Agip indicano che anche le energie propagandate come «pulite» sollevano problemi di accettazione da parte delle popolazioni interessate.

La vasta problematica ambientale in genere, e la valutazione dell'impatto ambientale delle fonti energetiche in particolare, dovrebbero ricevere in Italia la dovuta attenzione dalla classe politica, dai responsabili delle ricerche e dalla popolazione tutta. Il nostro paese è infatti caratterizzato da densità di popolazione elevata in media, e molto elevata nelle aree di pianura e lungo le coste. Proprio queste stesse aree sono le più appetibili, non solo per l'installazione di nuove centrali, ma anche per tutti gli altri usi, dall'agricoltura alle attività industriali, dal turismo agli insediamenti urbani.

In particolare nel nostro paese qualsiasi pianificazione di installazioni energetiche non può essere attuata se non vengono soddisfatte contemporaneamente due esigenze: le nuove installazioni, con relative ingombranti infrastrutture, dovranno essere sottoposte a una rigorosa procedura di pianificazione territoriale, con la definizione del conseguente impatto ambientale, e fatte oggetto di un serio sforzo tecnico e finanziario, teso a minimizzare tale impatto; le proposte di nuove installazioni debbono filtrare attraverso il vaglio responsabile delle popolazioni e delle amministrazioni locali interessate. E ciò esige un salto nella qualità e nei flussi globali di informazioni, gestiti con responsabilità scientifica e politica, da e verso le popolazioni e le amministrazioni interessate.

Scopo di questa nota è quello di sottoporre alla attenzione - e soprattutto alla critica costruttiva - di tutti gli interessati un approccio naturalistico e geochimico a tutta la problematica, con particolare riferimento alla contaminazione che è già stata - e che presumibilmente sarà indotta - dalle attività umane.

Questo tipo di impostazione tende a escludere da un lato la trattazione dei problemi di contaminazione derivanti dagli incidenti, sia perché in tal caso i flussi globali di contaminanti verso l'ambiente sono in genere molto più bassi di quelli che si osservano nel normale esercizio, e sia perché si tratta di argomento che per importanza e vastità richiederebbe una trattazione a se stante; d'altro lato anche i costi e l'impatto territoriale di ogni fonte energetica non vengono esaminati perché esulano dal taglio che si è voluto dare all'articolo e perché già esaurientemente dibattuti in altre pubblicazioni.

Definizione dei problemi e metodologia

Il punto di partenza di ogni discorso che verte sulle modificazioni antropogeniche dell'ambiente deve essere sempre costituito da un esame obiettivo, ed esprimibile in termini quantitativi, di quello che è lo stato dell'ambiente prima che intervengano le modificazioni dirette o indirette causate dall'uomo. In altri termini l'unico metro che deve essere usato è riferibile alle condizioni ambientali indisturbate, condizioni che non vanno considerate come qualcosa di statico e immutabile, ma come un divenire continuo, con fluttuazioni anche consistenti nello spazio e nel tempo, risultato questo di un gran numero di complessi equilibri dinamici tra i vari processi naturali.

Da quando, oltre due miliardi di anni fa, sono apparse sulla Terra le prime forme di vita, profonde e numerose sono state le modificazioni indotte dall'ambiente sul tipo di evoluzione delle specie viventi, e viceversa. Come variazione importante provocata dagli organismi viventi - in termini più generali, dalla biosfera - sull'ambiente, è significativo ricordare le profonde variazioni della composizione dell'atmosfera terrestre avvenute quasi un miliardo di anni fa (all'inizio o poco prima dell'era del Cambriano). L'atmosfera primordiale era molto ricca di anidride carbonica e di composti riducenti, mentre l'ossigeno era praticamente assente. L'esplosione dell'attività biologica, e segnatamente lo sviluppo della flora, hanno portato a gra-

duali combiamenti, rapidi in relazione ai tempi geologici, perché l'attività di fotosintesi clorofilliana è stata fortemente stimolata dalla elevata concentrazione di anidride carbonica (CO_2) nell'atmosfera. La elevata concentrazione di ossigeno, che si è così venuta a creare, ha modificato sostanzialmente l'ambiente inorganico, cambiando le caratteristiche chimico-fisiche sulla superficie della Terra e le modalità di circolazione di molti elementi, dal ferro che ossidandosi ha formato composti insolubili e non è stato più in grado di circolare, all'uranio che passando alla valenza + 6 è diventato un elemento solubile e mobile nell'ambiente superficiale.

Cambiamenti ancora più sostanziali si sono verificati nella evoluzione della biosfera, perché la presenza di ossigeno ha permesso il sorgere e l'evoluzione di forme di vita aerobica, di organismi cioè che vivono, si moltiplicano e si evolvono a spese di tutta una serie di reazioni biochimiche, il cui risultato finale è proprio la ritrasformazione dell'ossigeno in anidride carbonica. La composizione dell'atmosfera ha continuato a evolvere sino ai nostri tempi in base al raggiungimento di equilibri dinamici basati su azioni e controreazioni del tipo descritto, equilibri la cui estrema complessità è stata in parte chiarita solo dalle ricerche più recenti, e che comunque non possono essere costretti in schemi troppo semplificati senza perdere di veridicità.

Lo stesso tipo di discorso accennato per l'atmosfera terrestre vale per le altre «sfere geochimiche» oltre l'atmosfera, e cioè biosfera, idrosfera, litosfera e pedosfera. All'interno di ciascuna sfera geochimica, e tra le diverse sfere, si sono stabiliti degli equilibri dinamici che riguardano tanto il trasferimento di massa e di specie chimiche, quanto quello di energia, equilibri «faticosamente» raggiunti nei miliardi di anni di storia della Terra. In particolare, per quel che riguarda la parte più superficiale della Terra, all'interfaccia tra litosfera, biosfera, e atmosfera, si è stabilito un complesso equilibrio dinamico tra composizione dei suoli, dell'atmosfera, forme di vita, loro evoluzione e conseguente controreazione sull'ambiente; equilibrio che rappresenta un meccanismo di una fragilità estremamente elevata, a causa della stretta interdipendenza che lega strettamente tra loro tutti i processi che hanno luogo.

Così le forme di vita che si sono affermate si sono evolute anche in funzione della composizione dei suoli, dell'atmosfera e dell'idrosfera, interferendo in maniera sensibile, per mezzo della propria attività e relativi prodotti, sulla composizione di tali sfere. Quando l'uomo è apparso sulla Terra ha trovato un ambiente che aveva raggiunto un equilibrio sostanzialmente stabilizzato attraverso le vicissitudini di molte centinaia di milioni di anni, vicissitudini che sono state tratteggiate in precedenza molto succintamente. Negli ultimi milioni di anni l'ambiente nel quale si è sviluppata ed evoluta l'umanità ha mantenuto caratteristiche relativamente stabili sia per quanto riguarda la composizione dell'atmosfera, dell'idrosfera e della litosfera, e sia per le condizioni climatiche e per l'evoluzione geodinamica dei continenti.

La rapida crescita demografica, l'aumento dei consumi pro capite, il tumultuoso sviluppo delle attività umane in genere, e industriali in particolare, le aspettative maturate nelle popolazioni dei paesi emergenti - processi tutti che hanno indotto negli ultimi decenni un incremento di tipo esponenziale nelle richieste globali di energia e di materie prime - specie in epoca recente hanno fatto sentire pesanti effetti negativi sull'ambiente e pongono una serie di problemi drammatici per la conservazione dell'ambiente stesso e per la conservazione delle attuali specie viventi, ivi compresa quella umana.

Le considerazioni esposte in precedenza vogliono porre l'accento su un aspetto fondamentale, spesso trascurato, della problematica ambientale. Le condizioni ambientali attuali rappresentano il risultato di complessi equilibri dinamici che hanno raggiunto una relativa, sostanziale stabilità, dopo una tormentata storia di molte centinaia di milioni di anni. Questi equilibri presentano numerosi nodi critici, a partire dai quali è possibile innescare una serie di reazioni a catena - anche con cinetica di tipo esplosivo - capaci di modificare in maniera marcata e irreversibile le condizioni ambientali.

Equilibrio energetico della Terra e sue possibili alterazioni antropogeniche

Uno dei più importanti nodi critici è costituito dagli equilibri energetici tra il pianeta Terra e l'universo. La criticità del sistema dipende dal fatto che gli equilibri attuali dipendono fortemente dalla composizione dell'atmosfera, dalla qualità, quantità e distribuzione dei corpuscoli contenuti nell'atmosfera, dalle caratteristiche della superficie delle terre emerse e delle acque. Il bilancio generale energetico tra flussi in entrata e in uscita dal sistema terra sono influenzati anche da altri parametri, come ad esempio il calore sviluppato all'interno della Terra dagli elementi radioattivi naturali. La produzione di calore dei più importanti è la seguente: $^{238}U = 0{,}72$ calorie per grammo per anno; $^{235}U = 4{,}7$; $^{232}Th = 0{,}21$; $^{40}K = 0{,}21$. In particolare le percentuali di energia radiante solare che vengono direttamente o indirettamente trasformate in calore, utilizzate per il processo fotosintetico, trasformate in energia meccanica (venti, movimenti ondosi, correnti marine ecc.), utilizzate per mantenere in moto il ciclo dell'acqua, ecc., e infine irradiate di nuovo nello spazio, sono fortemente influenzate da ciascuna delle caratteristiche sopra citate.

Spostamenti anche limitati negli equilibri menzionati possono portare a variazioni irreversibili e sostanziali nelle condizioni climatiche e ambientali, ma non esistono ancora dati sicuri sulla valutazione quantitativa degli effetti provocati. Studi recenti hanno dimostrato che l'impiego sconsiderato di combustibili fossili e le attività industriali hanno provocato un aumento considerevole del tenore di CO_2 nell'atmosfera terrestre. Una piccola variazione nel tenore di CO_2, e/o delle particelle sospese nell'atmosfera, o di ozono nella parte più esterna dell'atmosfera, può fare aumentare la frazione di energia irradiata dal Sole che viene trattenuta nell'atmosfera terrestre sotto forma di calore, e l'aumento di temperatura conseguente può variare i rapporti tra quantità di acqua presente sul pianeta allo stato solido, allo stato liquido e allo stato di vapore, innescando un processo a catena le cui conseguenze non sono chiare neppure agli scienziati che si occupano dell'argomento. Comunque già il solo scioglimento di percentuali apprezzabili delle calotte polari, con il conseguente innalzamento di molti metri del livello degli oceani, porterebbe a conseguenze drammatiche.

Inoltre i granai del mondo (America settentrionale, Russia) sono stati sviluppati in aree che, anche per motivi morfologici, si prestavano a colture estensive, ma che presentano condizioni climatiche non ottimali e senza ampi limiti (piovosità scarsa, stagione calda non molto lunga ecc.). Variazioni anche limitate nelle condizioni climatiche potrebbero pertanto fare crollare la produttività di queste aree di sostentamento alimentare del mondo intero.

Se a scala globale l'energia prodotta dall'uomo è di ordini di grandezza inferiore al flusso di energia irradiata dal Sole sulla Terra (7×10^{16} chilocalorie all'anno contro circa 10^{21}) in molte aree a intensa urbanizzazione e/o industrializzazione la quantità di calore immessa dall'uomo nell'ambiente è già superiore al flusso di calore verso la superficie sviluppato all'interno della Terra, principalmente dalla radioattività naturale, che in totale è di circa $2{,}5 \times 10^{17}$ chilocalorie all'anno. Gli agglomerati urbani e industriali cresciuti in maniera abnorme e disordinata hanno già provocato variazioni apprezzabili nelle condizioni climatiche locali proprio perché, accanto alle quantità di energia prodotte e trasferite all'ambiente sotto forma di calore, hanno alterato gli equilibri preesistenti per quel che riguarda i rapporti tra flussi di energia in entrata e in uscita dal sistema. Un agglomerato di costruzioni si comporta infatti in maniera molto differente da un campo agricolo o una foresta per quel che riguarda le percentuali di energia irradiata dal Sole che vengono riflesse, diffuse, immagazzinate, trasformate e restituite all'ambiente.

Un esteso ricorso all'energia solare, come auspicato da più parti, deve confrontarsi anche con questo tipo di problematica. Un pannello solare si comporta infatti dal punto di vista degli equilibri energetici naturali in maniera completamente diversa da una fronda d'albero, perché il pannello tende a raccogliere tutta l'energia radiante che riceve e a trasformarla integralmente, in un tempo più o meno lungo, in solo calore e a trasferirla come tale all'ambiente. Le considerazioni sopra svolte non intendono portare alla conclusione drastica che l'energia solare è una risorsa che non deve essere presa in esame, ma vuole solo richiamare l'attenzione sull'esistenza di condizioni al contorno del sistema abiente, molto rigide, che debbono essere sempre tenute nel debito conto se si vuole evitare di adottare ipotesi di lavoro errate. In pratica il mantenimento della vita così come è configurata attualmente, e di tutti gli equilibri ambientali, presuppone necessariamente una ripartizione dei diversi utilizzi di tale energia (fotosintesi clorofilliana, riscaldamento dell'atmosfera, mantenimento del ciclo dell'acqua ecc.) in proporzioni non dissimili da quelle che si riscontrano attualmente. Ogni consistente manomissione di questi equilibri e di queste ripartizioni deve essere esaminata con estremo rigore, perché le conoscenze scientifiche sinora acquisite indicano che in tal caso si producono necessariamente alterazioni irreversibili; anche se non si è in grado di prevedere con precisione l'entità e i tempi di tali alterazioni.

Composizione delle sfere geochimiche e loro alterazioni antropogeniche

Già è stato ricordato l'aumento di concentrazione di CO_2 nell'atmosfera causato soprattutto dall'uso sempre più intenso di combustibili fossili. Tra le altre forme di inquinamento ambientale già osservate a scala locale o globale vanno ricordate l'eutrofizzazione, l'inquinamento da elementi stabili tossici e da composti organici tossici, le piogge acide.

L'eutrofizzazione consiste in una variazione marcata degli equilibri ecologici nei diversi tipi di corpi idrici (laghi, fiumi, mari) causata dall'immissione antropogenica di sostanze nutritive (soprattutto composti del fosforo) capace di innescare uno sviluppo parossistico di alcune specie vegetali, che con la loro attività e i loro prodotti di decadimento possono raggiungere il limite della eliminazione della vita aerobica per progressiva asfissia dell'ambiente.

Per quanto riguarda l'inquinamento da elementi stabili tossici si sono avuti casi gravi; dal dramma di Minamata Bay in Giappone con centinaia di intossicati e decine di morti per ingestione di pesce contaminato con mercurio dagli scarichi di industrie chimiche, ai più recenti 460 morti in Iraq per avvelenamento da mercurio contenuto nel pane, all'inquinamento delle acque naturali da metalli pesanti che è stato segnalato in tutte le aree del globo a intensa urbanizzazione e industrializzazione, alla distruzione dell'ecosistema del lago d'Orta e del lago di Varese da parte di scarichi industriali, al pericolo corso con il relitto della Cavtat con il suo carico di composti di piombo, all'inquinamento da arsenico dell'area di Manfredonia.

Ma l'aspetto più preoccupante è costituito dall'inquinamento che si è già verificato a scala globale sul nostro pianeta. Di recente sono stati condotti accurati studi

ELEMENTO	PRIMA DEL 1900	DOPO IL 1960
MANGANESE	0,145	0,139
ZINCO	0,077	0,224
PIOMBO	0,045	0,144
CADMIO	0,034	0,639

Tenore di elementi in traccia riscontrati in ghiacciai della Groenlandia. I valori riportati, espressi in microgrammi per chilogrammo di campione analizzato, sottolineano la gravità dell'inquinamento verificatosi in meno di un secolo. Solo il manganese si discosta da questo andamento.

Lo schema visualizza le numerose e profonde modificazioni indotte dall'ambiente sul tipo di evoluzione delle specie viventi. La cronologia è stata ottenuta da reperti geologici e paleontologici.

		CARBONE	PETROLIO	GAS	NUCLEARE FISSIONE	NUCLEARE FUSIONE	SOLARE FOTOVOLTAICA	METANOLO ED ETANOLO	GEOTERMIA
ESTRAZIONE MATERIE PRIME NECESSARIE	1	★	○	○	★	○	○	○	
	2	★	○	○	★	○	★	○	
	3	★	○	○	★	○	○	○	
	4	○	○	○	○	○	○	○	
	5	★★	★	★	★	○	★	★★	
	6	○	○	○	○	○	○	★	
	7	○	○	○	○	○	○	○	
	8	★★★	★	★	★★	★	★	★★	
	9	★★★	★	★	★	★	★	★★★★	
TRASPORTO STOCCAGGIO E TRASFORMAZIONE MATERIE PRIME	1	★★	○	○	○	○	○	○	
	2	★	○	○	★	○	○	○	
	3	★	○	○	★	○	○	○	
	4	○	★★	○	○	○	○	★★	
	5	★★	★★	○	○	○	○	★	
	6	★	★	○	○	○	○	○	
	7	○	★	○	○	○	○	○	
	8	★★★	★★	★	★★	○	○	★★	
	9	★	★	★	○	○	★	★★★	
COSTRUZIONE SMANTELLAMENTO IMPIANTI GENERAZIONE ENERGIA	1	★	○	○	○	○	○		
	2	○	○	○	○	○	★★		
	3	○	○	○	★★	★	○		
	4	○	○	○	○	○	○		
	5	★	★	○	★	★	★		
	6	○	○	○	○	○	○		
	7	○	○	○	○	○	○		
	8	★★	★	★	★★	★	★★		
	9	★	★	★	★	★	★★		
GENERAZIONE ENERGIA	1	★★★	★	★	○	○	○		○
	2	★★	★	○	○	★	○		★★★
	3	★	○	○	★★	★★	○		★★
	4	★	★★	★★★	○	○	○		★★
	5	★	★	★	★	★	★		★★★
	6	★★★	★★	★★	○	○	★★		★
	7	★★★	★★★★	★	○	○	○		★
	8	★★★	★★	★	★★	★★	★★★★		★★★
	9	★★	★	★	★	★	★★		★★
SISTEMAZIONE RIFIUTI	1	★★★	○	○	○	○	○	○	
	2	★★	★	○	○	○	★	○	
	3	★	○	○	★★	★	○	○	
	4	★	★	○	○	○	○	★	
	5	★★	○	○	★	○	○	○	
	6	○	○	○	○	○	○	○	
	7	○	○	○	○	○	○	○	
	8	★★★	○	○	★	○	○	○	
	9	★	○	○	★	★	○	○	

TIPO DI CONTAMINAZIONE E DI EFFETTI

1 - PARTICOLATO
2 - ELEMENTI STABILI TOSSICI
3 - ELEMENTI RADIOATTIVI
4 - COMPOSTI TOSSICI E CANCEROGENI
5 - INQUINAMENTO ACQUE E ALTERAZIONI IDROGEOLOGICHE
6 - ALTERAZIONI CLIMA
7 - PIOGGE ACIDE
8 - IMPATTO SUL TERRITORIO
9 - RISCHIO CONNESSO

ENTITÀ DEGLI EFFETTI

○ TRASCURABILE
★ DEBOLE
★★ CONSISTENTE
★★★ ELEVATO
★★★★ MOLTO ELEVATO

Schema riassuntivo dei vari tipi di contaminazione ambientale da parte delle più importanti fasi del ciclo produttivo delle diverse fonti energetiche. Nel caso del metanolo non sono state date specifiche sulle fasi di costruzione degli impianti e di generazione di energia perché essi sono in genere del tutto simili all'uso dei prodotti del petrolio per autotrazione, con il vantaggio di costituire una fonte d'inquinamento trascurabile perché non viene disperso il piombo, aggiunto come antidetonante, contenuto nelle benzine super e neppure prodotti organici tossici ciclici e policiclici. Nel caso della geotermia tutti gli effetti sono stati condensati nella fase di generazione di energia perché tutto il ciclo produttivo viene attuato per la massima parte nello stesso luogo. I dati riportati in questa tabella si riferiscono naturalmente alla tecnologia attualmente adottata. L'evoluzione della tecnologia degli impianti potrà contribuire a ridurre consistentemente le emissioni di prodotti tossici.

sul tenore di elementi tossici in traccia, in funzione del tempo di precipitazione della neve, nei ghiacciai della Groenlandia.

Indagini condotte su concrezioni di ossidi di ferro e manganese raccolte sul fondo del Mar Baltico e su carote dei sedimenti marini di fiordi della Norvegia, hanno fornito risultati del tutto analoghi; elementi fortemente tossici, quali Cu, Pb, Cd e Hg risultano arricchiti di 5-10 volte negli ultimi decenni.

Tutti i dati mostrano come negli ultimi decenni l'umanità ha causato variazioni consistenti nei tenori ambientali di elementi tossici, a scala globale; non si tratta di zone limitate della Terra, ma tutte le sfere geochimiche nel loro insieme sono interessate da questo processo di progressivo avvelenamento, dalle acque oceaniche, all'atmosfera, alla biosfera nel suo complesso.

Per quanto riguarda i composti organici tossici, a parte i disastri ecologici che hanno interessato zone limitate, come la dispersione di diossina a Seveso, alcuni composti tossici artificiali (ad esempio il DDT) mostrano attualmente tenori apprezzabili anche in zone molto remote e presentano in parte la caratteristica negativa di accumularsi lungo la catena alimentare. L'aspetto più inquietante del problema è costituito dalla ignoranza quasi totale sul livello di contaminazione ambientale da parte delle numerosissime sostanze organiche, prodotte e disperse dall'umanità, e sulla tossicità di tali sostanze per esposizioni prolungate a dosi basse o molto basse.

Si sa ad esempio che estese aree del nostro, come di altri paesi, sono contaminate dalla presenza di solventi clorurati, reperibili in quantità apprezzabili persino nelle acque potabili (è il caso, ad esempio, della Pianura Padana), ma non si sa quasi nulla sulla distribuzione dei pesticidi, dei composti fenolici, benzenici e policiclici immessi nell'ambiente dall'industria chimica e dall'uso dei combustibili fossili. Si sa che moltissimi di questi composti sono tossici e numerosi sono cancerogeni, ma praticamente nulla si sa circa i rischi connessi a lunghe esposizioni a basse dosi.

Negli ultimi anni, specie nei paesi industrializzati, si è assistito anche a un aumento dell'acidità delle acque piovane. Questo fenomeno trae origine dai composti dello zolfo, e subordinatamente dell'azoto, che vengono immessi nell'atmosfera dall'uso dei combustibili fossili. I prodotti dello zolfo e dell'azoto, solitamente indicati con la formula SO_x e NO_x perché costituiti da miscugli di ossidi a diverso stato di ossidazione, vengono alla fine trasformati dall'ossigeno dell'aria e dall'acqua in acido solforico e acido nitroso e nitrico.

Il fenomeno delle piogge acide ha assunto aspetti molto preoccupanti perché è frequente avere piogge con pH inferiore a 4, e questi valori hanno un effetto negativo non solo sulle costruzioni e sulle opere d'arte (spesso di marmo), ma anche sull'agricoltura e sulla qualità delle acque naturali.

I diversi tipi di inquinamento ambientale presentano tutti una drammatica caratteristica: la estrema rapidità dei cambiamenti della composizione dell'ambiente nella sua globalità in rapporto ai tempi lunghissimi che sono stati necessari per il raggiungimento degli equilibri, che erano ormai sostanzialmente stabilizzati quando l'uomo ha fatto la sua comparsa sul nostro pianeta. E l'ignoranza sugli effetti a lungo termine di questi cambiamenti è drammaticamente ampia e profonda (si veda l'articolo *Pioggia acida* di G. E. Likens, R. F. Wright, J. N. Galloway e T. J. Butler in «Le Scienze», n. 136, dicembre 1979).

Radioattività naturale

Tra gli elementi tossici figurano anche elementi radioattivi. Di questi alcuni sono di origine naturale e altri sono stati prodotti artificialmente dall'uomo. Tra quelli naturali di particolare importanza per la loro tossicità sono il rado, il radio e l'uranio. È stato anche ricordato che gran parte del calore prodotto all'interno della Terra è originato dalla radioattività degli elementi della famiglia dell'uranio e del torio e dal potassio 40.

È opportuno sottolineare che la radioattività, o per meglio dire tutta la serie di radiazioni che vanno dalle particelle alfa alle beta, dalle gamma sino alle radiofrequenze passando attraverso lo spettro del visibile, hanno sempre fatto parte integrante dell'ambiente terrestre. Nel corso della lunga evoluzione geologica della Terra si è avuta una progressiva diminuzione della radioattività presente sulla superficie terrestre, sia perché è aumentato l'effetto schermante dell'atmosfera e sia perché gli elementi radioattivi, proprio perché tali, hanno diminuito progressivamente la loro concentrazione, e quindi l'intensità delle radiazioni emesse. È interessante ricordare che tanto la vita sulla terraferma come pure il rapido aumento del fitoplancton nelle acque superficiali si sono potuti realizzare, circa un miliardo di anni fa, solo quando la concentrazione di ozono nell'alta atmosfera è stata sufficiente a schermare la gran parte delle radiazioni ultraviolette provenienti dal Sole.

Sicuramente la radioattività naturale rientra fra le condizioni ambientali che hanno influito sulla evoluzione biologica, in maniera analoga a quanto avvenuto per la composizione chimica delle diverse sfere geochimiche; ma su questo importante problema non si dispone di alcuna informazione sicura. Anche in questo caso sono stati raggiunti aggiustamenti tra intensità di radiazioni nelle diverse aree ed evoluzione delle specie biologiche presenti. Tuttavia non si ha alcuna idea del ruolo diretto che ha svolto la radioattività ambientale naturale sulla direzione e sulla velocità della evoluzione biologica, attraverso il ben noto effetto delle radiazioni sul materiale genetico, cioè attraverso la induzione di mutazioni nuove.

Allo stato attuale delle cose l'inquinamento radioattivo da parte dell'uomo è contenuto in livelli molto bassi. In pratica le contaminazioni a scala globale già verificatesi sono dovute quasi esclusivamente alle attività militari.

Che la radioattività faccia parte integrante dell'ambiente naturale ha trovato una conferma nella scoperta dei resti di un reattore naturale che è diventato critico molto tempo fa nel Gabon.

Circa 1,8 miliardi di anni fa nella zona di Oklo, in corrispondenza di un giacimento di uranio con tenori molto elevati (sino al 70 per cento di uranio in alcune zone del giacimento), e a seguito del verificarsi di tutta una serie di circostanze ambientali favorevoli, ha raggiunto lo stato di criticità un reattore nucleare del tutto naturale. Questo reattore naturale ha funzionato, probabilmente con numerose interruzioni e con fluttuazioni consistenti nella potenza, per un periodo lunghissimo, e cioè circa 500 000 anni. Che un giacimento di uranio abbia potuto dare luogo, in condizioni del tutto naturali, a reazioni nucleari del tutto simili a quelle prodotte artificialmente dall'uomo, è una conseguenza della circostanza che 1,8 miliardi di anni fa il tenore dell'isotopo ^{235}U rappresentava il 3,16 per cento dell'uranio, proprio cioè la percentuale che si ottiene nell'uranio arricchito impiegato attualmente nei reattori artificiali (si veda l'articolo *Un reattore naturale a fissione* di G. A. Cowan in «Le Scienze», n. 99, novembre 1976).

Il progressivo impoverimento in natura dell'^{235}U in funzione del tempo dipende dal valore del suo periodo di dimezzamento che è molto più breve di quello dell'^{238}U, rispettivamente di $7,13 \times 10^8$ contro $4,51 \times 10^9$ anni.

È stato possibile studiare esaurientemente il comportamento in natura di molti degli elementi radioattivi che si sono originati a seguito delle reazioni nucleari naturali. Il risultato invero inatteso di tali studi geochimici è che durante e dopo il funzionamento del reattore naturale, e cioè per quasi due miliardi di anni, non si sono verificate migrazioni di rilievo dei diversi elementi radioattivi e stabili presenti nella zona del reattore. In particolare tutti gli elementi sono rimasti nel punto in cui erano o hanno subito migrazioni limitatissime (dell'ordine dei centimetri o dei metri), tranne gli elementi alcalini (Li, Na, K, Rb e Cs), il Mo, il Cd, il Pb, il Ba e i gas. In pratica la quasi totalità degli elementi contenuti nelle «scorie radioattive», ivi compresi Pu, Sr, Am ecc., sono rimasti sul posto, e dopo il lunghissimo tempo trascorso dalla loro formazione, si sono trasformati quasi integralmente in elementi stabili.

I risultati sopra riportati sono di indubbia utilità negli studi finalizzati alla sistemazione definitiva dei residui radioattivi, perché non è difficile reperire zone in cui sussistono condizioni ambientali del tutto simili a quelle che si sono avute nella zona di Oklo per quasi due miliardi di anni. In condizioni ambientali di questo tipo si può nutrire la ragionevole certezza che anche residui radioattivi, non contenuti in recipienti perfettamente chiusi, resterebbero immobilizzati fino alla loro trasformazione completa in elementi stabili.

SORGENTI DI ESPOSIZIONE	DOSE GLOBALE MEDIA (MILLIREM)
ESPOSIZIONE ANNUA ALLA RADIOATTIVITÀ AMBIENTALE	365
VIAGGI AEREI	0,4
PRODUZIONE FOSFATI	0,04
PRODUZIONE ATTUALE ANNUA GLOBALE DI ENERGIA ELETTRICA DAL CARBONE (10^6MWe)	0,02
ESPOSIZIONE AI NORMALI PRODOTTI DI CONSUMO	3,0
PRODUZIONE ATTUALE ANNUA DI ENERGIA ELETTRONUCLEARE (10^5 MWe)	0,6
FALL OUT DA ESPLOSIONI NUCLEARI, VALORE MEDIO DEL PERIODO 1951-1976	30
RADIOGRAFIE E RADIOSCOPIE	70

Dosi medie all'uomo delle diverse sorgenti di radiazioni più importanti. Nell'anno 2000 l'attuazione dei programmi di installazioni nucleari (2×10^6MWe) prevede una esposizione globale equivalente a 15 giorni di radiazione naturale. Ricerche più recenti hanno dimostrato che la produzione di energia da carbone viene sottostimata in tale tabella. La dose media globale annua (espressa in giorni equivalenti di esposizione) dovuta a radioattività naturale in Italia è in genere compresa tra 100 e 200 millirem. In alcune aree del Lazio e della Campania si hanno valori molto più elevati.

Impatto ambientale delle diverse fonti energetiche

La produzione di energia si è andata sempre più configurando come una produzione concentrata in impianti relativamente poco numerosi, ma di potenza unitaria sempre più elevata. Questa tendenza è stata alimentata da due classi di «esigenze». La prima è riconducibile all'abbassamento del costo di produzione unitario con economie di scala, e la seconda esigenza è riconducibile al «desiderio» dell'utente di premere un pulsante e avere a disposizione tutta l'energia che ritiene necessaria, senza preoccuparsi dei fastidi e degli inconvenienti connessi alla sua produzione.

Le attuali gigantesche centrali di produzione di energia elettrica, di qualsiasi tipo esse siano, provocano consistenti e palesi alterazioni ambientali, per la semplice circostanza che in uno spazio limitato vengono concentrati flussi artificiali terribilmente elevati di materia e di energia. Per basilari principi termodinamici, che non possono essere mutati dalla tecnologia, la parte prevalente dei flussi di energia in gioco viene trasformata in calore. Il problema dell'inquinamento termico dell'ambiente sussiste infatti, sostanzialmente con la stessa gravità, per tutti i tipi di centrali di produzione, e lo smaltimento del calore nell'atmosfera invece che nelle acque naturali non rappresenta una soluzione definitiva del problema per i motivi precedentemente discussi. Una attenuazione dell'impatto ambientale può essere ricercata soprattutto nelle tecnologie di cogenerazione, che permettono l'utilizzazione diretta, ad esempio per il riscaldamento domestico, del calore prodotto dalle centrali.

Inoltre anche una parte apprezzabile degli elevatissimi flussi di materia esistenti nelle centrali viene necessariamente dispersa nell'ambiente. Tali flussi di sostanze, anche tossiche, verso l'ambiente potrebbero essere ridotti praticamente a zero dalla moderna tecnologia, ma con costi che praticamente raggiungerebbero valori infiniti. Per i motivi esposti, forse troppo in sintesi, in precedenza, non è possibile parlare di fonti di energia pulite contrapposte ad altre fonti altamente inquinanti.

L'unico discorso costruttivo può essere impostato su un raffronto tra l'entità delle perturbazioni ambientali causate dalle diverse fonti energetiche. Questo raffronto deve essere completo, deve cioè comprendere tutte le fasi verticali del complesso di operazioni che portano al risultato finale di produzione della stessa quantità e qualità di energia, se si vuole che le conclusioni finali poggino su basi scientifiche. È questo un importante aspetto del problema che molto spesso viene del tutto trascurato, o volutamente sottostimato, e richiede pertanto qualche approfondimento.

In questa sede verrà fornito un quadro complessivo, anche se necessariamente lacunoso e incompleto, della contaminazione ambientale causata dalle diverse fonti energetiche.

La tabella di pagina 100 presenta i risultati di una vasta analisi volta a fornire una informativa omogenea sulla contaminazione ambientale provocata dalle più importanti fonti energetiche, considerata ciascuna nelle diverse fasi del ciclo produttivo. I dati riportati in maniera sintetica nella tabella vogliono rappresentare solo un punto di riferimento e una base di partenza per l'approfondimento della vastissima problematica considerata. È difficile infatti presentare un quadro sintetico di questa ampiezza senza necessariamente incorrere in inesattezze.

La presentazione di questa complessa analisi trova giustificazione solo nella necessità di mettere un po' d'ordine e di fornire spunti di riflessione per una corretta impostazione della pianificazione energetica, che tenga conto anche degli effetti sull'ambiente e sulla salute delle popolazioni interessate. I dati riportati necessitano di qualche commento tendente a chiarire i termini reali del problema e a sottolineare l'importanza di alcuni aspetti che vengono in genere taciuti o sottovalutati nel dibattito in corso nel nostro paese. Così, ad esempio, la energia geotermica, la nucleare da fusione e la solare non sono energie così «pulite» come spesso viene fatto credere.

L'energia geotermica ha visto crescere i suoi sostenitori anche perché ritenuta una energia pulita e con impatto sull'uso del territorio del tutto trascurabile. In effetti anche questo tipo di energia rinnovabile provoca effetti apprezzabili. Va ricordato infatti che in Italia ben due campi geotermici non sono ancora stati messi in produzione anche per i problemi ambientali sollevati. Il primo caso riguarda il campo geotermico di Cesano, presso Roma. I fluidi rinvenuti hanno una temperatura molto elevata, ma non è stato possibile sinora effettuare prove complete di produzione perché si tratta di un fluido con salinità eccezionalmente elevata (più di 300 grammi/litro) che contiene per giunta tenori molto alti di elementi tossici quali il boro (10 grammi/litro) e l'arsenico, e la dispersione in superficie di quantità consistenti di questi fluidi avrebbe causato un danno ambientale troppo elevato. Il secondo riguarda il campo geotermico dei Campi Flegrei dove l'elevata densità di popolazione e i timori per i possibili effetti sulla salute hanno reso molto difficile l'ubicazione dei pozzi produttivi, prima, e rallentato, poi, le prove di produzione.

I fluidi geotermici che vengono portati in superficie per produrre energia hanno tenori eccezionalmente elevati di elementi e composti tossici (B, Hg, As, CO_2, H_2S). Parte di questi tossici, e specie quelli volatili, vengono comunque dispersi nell'ambiente anche se viene prevista la reiniezione in profondità dei fluidi utilizzati. E la reiniezione può provocare movimenti del suolo e attività microsismica. Anche valori molto elevati di radioattività vengono in genere riscontrati nei fluidi geotermici. Tale radioattività è dovuta soprattutto al ^{222}Rn. Si pensi che il solo campo geotermico di Larderello immette nell'atmosfera circa 3000 curie per anno di Rn. E questo con una potenza installata di circa 400 MWe.

In termini più generali l'energia geotermica rappresenta un caso emblematico per sottolineare l'importanza del contesto ambientale e socioeconomico, nel quale essa viene sviluppata e inserita, nel determinare l'entità degli effetti sull'ambiente e la salute pubblica. Nell'area di Larderello la finalizzazione primaria di decine di chilometri quadrati è quella di produrre energia geotermica, ma non sussistono gravi problemi perché l'energia geotermica ha costituito il punto di riferimento per lo sviluppo di una ampia area. Molto diverso è il caso di campi geotermici che vengano rinvenuti in aree con elevata densità di popolazione. La entità della perturbazione ambientale, la necessità di superficie disponibile (1 chilometro quadrato circa per ogni 30 MWe installati) possono ostacolare o al limite sconsigliarne l'impiego. È invece da sottolineare non solo la convenienza, ma anche il trascurabile impatto ambientale, dell'impiego di fluidi a bassa entalpia (temperatura $\leq 100°C$) per l'utilizzazione diretta nel riscaldamento di abitazioni e serre o per usi industriali purché sia adottata la reiniezione dei fluidi.

Per quel che riguarda l'energia da fusione nucleare non si dispone di dati sicuri di riferimento perché la tecnologia è ancora in fase di sviluppo, ma risulta comunque ben chiaro che anche questo tipo di energia provocherà necessariamente consistenti flussi di sostanze tossiche sia radioattive che stabili verso l'ambiente. Le conoscenze attuali indicano infatti che un reattore a fusione di circa 500 MWe causerebbe in primo luogo un rilascio nell'ambiente di 2,6 curie al giorno di trizio, e il trizio è un isotopo dell'idrogeno che entra facilmente nella catena alimentare. Si avrebbero inoltre contaminazioni ambientali da parte dei metalli tossici necessari alla costruzione degli schermi (vanadio, berillio, ecc.).

L'esercizio di un tale reattore porterebbe inoltre alla formazione di elementi radioattivi a seguito dell'attivazione degli elementi stabili che costituiscono le strutture (Zr, Mo, Cr, Co, Mn). Lo smantellamento di un reattore a fusione e la sistemazione delle parti radioattive porrebbe cioè problemi del tutto simili a quelli incontrati nel caso dei reattori a fissione.

Un'altra fonte energetica che merita qualche commento supplementare è il carbone. Questo combustibile fossile, le cui riserve accertate sono talmente elevate da garantire un approvvigionamento energetico all'umanità per centinaia di anni, subirà un crescente impiego in tutti i paesi industrializzati, con lo scopo di sostituire gradualmente il petrolio nella generazione di energia. La problematica sollevata da un massiccio ricorso al carbone è davvero vasta e complessa perché spazia dal reperimento degli ingentissimi mezzi finanziari necessari a lanciare e sostenere l'impresa carbone, alla pianificazione di tutte le opere strutturali e infrastrutturali, al supporto tecnologico richiesto per ottimizzare il rapporto costi-benefici e per minimizzare le emissioni e quindi i danni all'ambiente, alle ricerche ambientali necessarie per comprendere quali sono i tossici più pericolosi emessi e quali le loro vie critiche verso la biosfera e l'uomo in particolare.

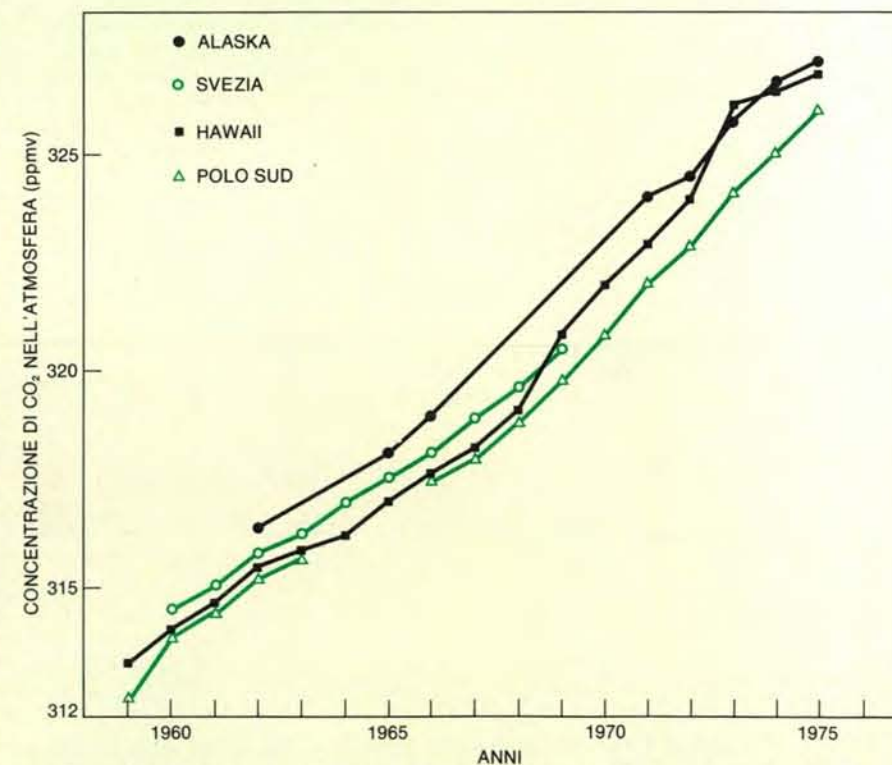
Negli ultimi anni è apparso un numero molto elevato di articoli tecnici che forniscono nuovi dati e sollevano nuovi problemi relativi all'impatto ambientale di questa fonte energetica; in particolare alcuni documenti molto recenti esaminano compiutamente l'ampia problematica derivante da un esteso ricorso al carbone. In questa sede vengono esaminati solo i più importanti problemi ambientali che saranno posti nella peculiare situazione italiana, sia ambientale che socioeconomica, da un massiccio ricorso all'impiego del carbone.

Al fine di non penalizzare troppo l'uso del carbone è opportuno impostare la relativa problematica ambientale in rapporto all'uso del petrolio e del nucleare. Queste due forme di energia rappresentano infatti dei punti di riferimento obbligati perché la prima tende a essere gradualmente sostituita dal carbone, a partire dai grandi impianti, e la seconda risulta essere la concorrente del carbone per le nuove installazioni energetiche.

Già il solo esame dei dati sintetici, presentati nella tabella di pagina 100 e che si riferiscono alla tecnologia attualmente in uso, indica che il carbone può provocare in alcune fasi del ciclo energetico una contaminazione ambientale più consistente di quella causata dalle fonti energetiche prese a confronto. Se si considera inoltre che nell'attuale momento tecnologico e socioeconomico è previsto il ricorso a centrali a carbone con potenza di circa 2500 MWe (4 gruppi da 600-650 MWe) e si fa riferimento ai flussi di materia e di energia in gioco, se ne ricava che la contaminazione ambientale risulta essere consistente. Così per una sola centrale di questo tipo occorrono circa 6 milioni di tonnellate di carbone per anno in entrata e quasi 1 milione di tonnellate di materiale di risulta in uscita (ceneri, scorie, residui dei trattamenti ecc.). Una simile centrale produce 100 000 tonnellate equivalenti per anno di anidride solforosa che, pur nel rispetto della normativa vigente, sarebbero immesse integralmente nell'ambiente.

Ma l'impatto ambientale forse più consistente nel medio e nel lungo periodo è costituito dal particolato trascinato dai fumi delle ciminiere. Purtroppo non si dispone di dati analitici completi sulla composizione del particolato. Si sa soltanto che alcuni elementi tossici che formano composti volatili (Hg, Se, As, Cd, Sb, Pb) passano direttamente nell'atmosfera allo stato di vapore o sono contenuti nelle frazioni più sottili (inferiori a 1 micrometro) che non sono fermati dai sistemi di abbattimento in uso.

Sussiste inoltre il problema della sistemazione delle ceneri e delle scorie prodotte dalla combustione del carbone. La sistemazione e/o l'utilizzazione di 5-7 milioni di tonnellate di ceneri per anno, previste per la fine del secolo, porrà una serie di quesiti che non saranno risolvibili con grande facilità e nel breve termine, considerando anche le dimensioni territoriali ridotte del nostro paese.

In effetti l'impresa carbone riveste un

Variazione, indotta dall'uomo, del tenore di anidride carbonica nell'atmosfera. I risultati dei quattro studi diversi, effettuati in zone differenti della Terra, indicano tutti una variazione consistente.

ruolo determinante per lo sviluppo energetico del nostro, come di altri paesi. Particolarmente per l'Italia è assolutamente necessario che il problema della contaminazione conseguente all'impiego di grandi quantità di carbone venga affrontato, mutuando dall'energia nucleare lo sforzo tecnologico - con relativo supporto finanziario - tendente a minimizzare le emissioni di sostanze tossiche, e che sia dato il via alla programmazione di estese ricerche ambientali e tossicologiche tendenti a chiarire le modalità di circolazione di questi tossici.

La storia dell'impiego del petrolio e dei suoi derivati può d'altra parte insegnare che è necessario, da un lato, affrontare con il dovuto impegno il problema della ubicazione delle nuove installazioni (potendo mutuare anche in questo caso dall'esperienza maturata in campo nucleare dallo studio dei siti) e, dall'altro lato, aggiornare e modificare la normativa in funzione delle acquisizioni tecnologiche e degli studi ambientali.

Questa strada costituisce l'unico percorso agibile per minimizzare i danni ambientali e ottenere quindi il necessario consenso alle nuove installazioni energetiche a carbone da parte delle popolazioni e delle amministrazioni interessate.

Rischio connesso

La definizione di rischio, come prodotto della probabilità che si verifichi un evento dannoso per la entità del danno causato da tale evento, rappresenta un tentativo di razionalizzazione della materia, perché si presta a essere quantificata per qualsiasi tipo di evento. Non è certo comunque che il rischio così definito sia l'entità più adatta alla comprensione completa di un problema in se stesso molto complicato. Così, ad esempio, lo stesso valore di rischio relativo a due eventi che si verificano con probabilità molto diversa tra loro viene recepito e vissuto dalla gran parte delle persone in maniera molto diversa. Già questa semplice osservazione indica che in effetti tale definizione di rischio non è la più adatta a fornire confronti validi perché il suo significato viene inficiato da reazioni umane emotive. Se infatti si tenesse, ad esempio, nel debito conto il rischio connesso alla circolazione automobilistica (circa 10 000 morti per anno nel nostro paese, e causa di morte più frequente per persone al di sotto dei 35 anni), l'impiego dell'auto sarebbe limitato all'indispensabile.

La prima osservazione basilare riguarda l'utopia del rischio zero. A qualsiasi attività umana - ivi comprese le azioni più normali e quotidiane - corrisponde un certo valore di rischio. Il rischio è un qualcosa di intimamente connaturato all'uomo e molte decisioni vengono quotidianamente prese anche in base alla valutazione dei rischi che comportano le diverse possibili scelte.

Per quanto riguarda il rischio connesso alla produzione di energia, come in qualsiasi altro tipo di attività industriale, è bene - per maggiore chiarezza - tenere separato il rischio corso dagli addetti ai lavori, da quello che incombe sulle popolazioni all'esterno degli impianti, o comunque estranee al processo produttivo. I rischi degli addetti ai lavori, malgrado vi siano ancora numerose incertezze nei calcoli attuali, sono valutabili con una approssimazione sufficiente almeno a fornire una scala di valori fortemente indicativa.

I rischi degli addetti ai lavori relativi all'utilizzo di diverse fonti energetiche sono comprensivi di tutte le operazioni necessarie alla produzione, dall'attività mineraria alla sistemazione finale dei rifiuti. I risultati riportati sono in parte sorprendenti perché, ad esempio, una fonte di energia rinnovabile quale la produzione di metanolo, sostenuta da molti «ecologi», presenta un rischio per gli addetti ai lavori di ordini di grandezza più elevati di altre fonti. In pratica il rischio è condizionato in primo luogo dalle quantità di materiali, maneggiati in tutte le fasi delle lavorazioni, necessarie per unità di energia prodotta.

I bassi rischi connessi all'impiego di gas naturale e all'energia nucleare sono legati anche alla limitata quantità di materiali che debbono essere direttamente maneggiati per unità di energia prodotta. Le stime disponibili del rischio per la popolazione riguardano solo gli effetti diretti.

Ma per una valutazione dei rischi globali cui è sottoposta la popolazione non si dispone dei dati scientifici indispensabili per trarre conclusioni univoche. Questa incapacità discende dalle gravi lacune conoscitive sui danni provocati all'uomo da una esposizione a dosi molto basse per un periodo molto lungo. Ciò vale per tutte le fonti di inquinamento, ma particolarmente per quelle non nucleari.

Sono già stati discussi i risultati di studi molto recenti che mostrano in maniera incontrovertibile che esiste un inquinamento antropogenico a scala globale di molti elementi e composti tossici; e nel prossimo futuro è previsto un aumento dei flussi antropogenici di questi elementi verso l'ambiente. Molti di questi elementi presentano inoltre la caratteristica negativa di tendere ad arricchirsi progressivamente lungo la catena alimentare.

L'aspetto più drammatico è costituito dalla circostanza che non esiste attualmente la capacità scientifica necessaria per fornire stime attendibili sui danni diretti e indiretti che saranno provocati all'umanità da queste sostanziali alterazioni dell'ambiente. Sono stati discussi i risultati di studi molto recenti che indicano le variazioni già osservate nella composizione dell'atmosfera (aumento di CO_2 e di pulviscolo) e previsioni circa la variazione delle condizioni climatiche della Terra. Chi è attualmente in grado di fornire dati, non diciamo attendibili, ma solo largamente indicativi su quali e quanti sarebbero i danni diretti o indiretti alle popolazioni, conseguenti a una variazione della temperatura media di 1, 2 o 5 gradi centrigradi?

Rischio espresso in persone per giorno rese inabili per ogni megawatt annuo di energia prodotta e riferito all'intera durata dell'impianto, e cioè circa 30 anni. Gli istogrammi sono proporzionali al numero di persone rese inabili per giorno, numero riportato per ogni tipo di energia. Queste stime sono indicative e servono solo ai fini di un raffronto tra gli effetti delle diverse fonti energetiche.

Ciclo geochimico del piombo. Tutte le grandezze sono espresse in 10^9 grammi di piombo; tutti i flussi sono espressi per anno. Risalta con evidenza la grave alterazione del ciclo da attività umane. Così contro 18×10^9 grammi immessi nell'atmosfera da processi naturali ben 440×10^9 grammi sono immessi da attività umane. In effetti la situazione è più grave di quello che appare dal diagramma perché gran parte degli altri 4000×10^9 grammi di piombo estratti annualmente viene direttamente o indirettamente dispersa nell'ambiente. Molti valori riportati sono solo indicativi: quasi sicuramente è sovrastimata la quantità di piombo presente nelle acque continentali e sotterranee. Un semplice calcolo indica che il tempo di residenza medio del piombo nell'atmosfera è di 15 giorni circa, mentre nell'idrosfera (oceani) è di centinaia di anni.

Conclusioni

Nelle pagine precedenti è stato compiuto il tentativo di impostare un discorso scientificamente valido sulla strada, molto lunga e accidentata anche per gli addetti ai lavori, della valutazione della contaminazione ambientale da parte delle diverse fonti energetiche.

Il problema fondamentale al quale si riferiscono, direttamente o indirettamente, tutti i quesiti e tutte le preoccupazioni delle popolazioni interessate a nuove installazioni industriali, e in particolare energetiche, è individuabile nelle correlazioni esistenti tra inquinamento ambientale e possibili danni alla salute dell'uomo. Tanto in Italia quanto in altri paesi molte comunità locali e associazioni ecologiche hanno infatti intrapreso campagne contro le nuove installazioni energetiche, sia convenzionali sia nucleari, per evitare l'inquinamento del loro territorio. In effetti il vero movente - a volte palesemente enunciato e a volte implicito nelle richieste -, che è alla base di questo tipo di contestazione, non è tanto e solo da ricercare in un interesse a salvaguardare il paesaggio e l'ambiente nel suo complesso, ma è piuttosto individuabile nel timore, non del tutto ingiustificato, che l'inquinamento ambientale causato dalle centrali provochi, nel medio e nel lungo periodo, danni alla salute delle popolazioni. A tale quesito legittimo e fondamentale la scienza moderna non è purtroppo capace di fornire risposte complete e soddisfacenti. Questa incapacità deriva in primo luogo dalla carenza di informazioni attendibili sulle modalità di circolazione delle sostanze ed elementi tossici nell'ambiente sino a interessare la biosfera in generale, e l'uomo in particolare. In secondo luogo mancano quasi del tutto informazioni complete e attendibili sui danni provocati agli animali, e in particolare all'uomo, da esposizioni molto prolungate a dosi o concentrazioni molto basse di sostanze tossiche.

Questa ignoranza è particolarmente vasta e profonda per quel che riguarda i rischi derivanti nel lungo termine da tenori di elementi non radioattivi e composti organici tossici, anche molto bassi in assoluto, ma anomali rispetto al valore di fondo naturale, nell'ambiente e nella catena alimentare. L'impiego pacifico dell'energia nucleare ha costituito una vera rivoluzione dell'era industriale perché è stato adottato, per la prima volta nella storia dell'umanità, il principio di tendere alla riduzione dei rischi, cui vengono necessariamente sottoposti i lavoratori e le popolazioni, a valori prossimi allo zero. Questo tipo di metodologia, che ha profonde radici nell'aspirazione a una palingenesi dopo gli orrori e le stragi terrificanti causate dall'impiego bellico dell'energia nucleare, ha portato a una impostazione filosofica del tutto nuova. Non è stato cioè accettato il rischio «oggettivamente insito nel tipo di lavorazione», come sempre è stato fatto, e viene tuttora fatto negli altri tipi di attività umane, dall'attività edilizia alle lavorazioni meccaniche, dai trasporti all'attività mineraria ed estrattiva in genere. In tutti questi settori gli addetti ai lavori sanno benissimo che sotto il profilo del calcolo delle probabilità si ha un morto per ogni tanti metri cubi di costruzioni ultimate, per tante tonnellate di minerale estratto, per tante tonnellate per chilometro di merci trasportate ecc. E questi dati vengono accettati come una cosa normale e «necessaria al progresso dell'umanità». Né una impostazione dissimile è riscontrabile nei paesi politicamente diversi. Tale impostazione è cioè «accettata» sia nei paesi socialisti come nei paesi dell'emisfero occidentale.

Nel settore nucleare è stato per la prima volta compiuto il tentativo di impostare in maniera razionale e sistematica la vasta problematica connessa ai possibili incidenti anche gravi causati direttamente o indirettamente da eventi naturali catastrofici e rari. Lo studio dei siti, ad esempio, è assurta a scienza in tale ambito con lo scopo di giungere alla eliminazione, o quanto meno alla riduzione, degli effetti

Schema semplificativo delle alterazioni ambientali provocate dalle diverse fonti energetiche e relativi effetti negativi. È necessario attuare una sorveglianza continua ed estesa a tutta la superficie del pianeta dello stato di salute dell'ambiente onde evitare che vengano superati dei valori di soglia di particolari condizioni, oltre i quali si possono scatenare dei processi a catena con effetti molto gravi e irreversibili.

sugli impianti di eventi eccezionali come i terremoti. La procedura di effettuare studi completi sulla situazione sismotettonica del sito destinato a ospitare una centrale nucleare, con lo scopo di imporre una accelerazione di progetto corrispondente al terremoto massimo possibile cui debbono resistere tutti gli impianti senza compromettere la sicurezza dei lavoratori e delle popolazioni interessate, rappresenta ad esempio un tipo di metodologia che dovrebbe essere seguita, e che sicuramente sarà seguita anche per installazioni importanti di altro tipo.

Anche per quel che riguarda le relazioni tra basse dosi di radiazioni ed effetti sulla salute, gli studi degli ultimi quarant'anni hanno portato a conoscenze molto più avanzate, anche se non ancora sufficientemente dettagliate e complete, rispetto a quelle disponibili nel settore della tossicologia dei metalli pesanti e dei tossici organici per esposizioni basse o molto basse.

La definizione e la misura della dose sono il punto di riferimento essenziale per articolare qualsivoglia discorso radioprotezionistico. È infatti la dose di radiazione assorbita dalle forme viventi - e quindi anche dall'uomo - che condiziona il tipo e l'entità del danno probabile e quindi del rischio. È stato possibile stabilire con sufficiente approssimazione quali sono le dosi al di sopra delle quali si ha una probabilità consistente di avere un conseguente effetto dannoso. Queste dosi sono relativamente elevate e debbono interessare tutto il corpo umano o suoi organi importanti. Il discorso diventa più complesso per le dosi molto più basse, le quali non provocano con sicurezza alcun effetto immediato, ma si assume possano provocare effetti anche gravi nel medio e nel lungo termine.

I possibili danni da dosi piccole o molto piccole sono di natura probabilistica nel senso che possono insorgere, ma non si verificano necessariamente su un singolo individuo esposto. Anche se le probabilità che l'evento dannoso si verifichi sono estremamente basse, il rischio conseguente è apprezzabile perché i danni tardivi prodotti dalle radiazioni comprendono malattie molto gravi come la leucemia e i tumori. Le stime più recenti, basate sullo studio accurato dei pochi casi che permettevano di trarre conclusioni con ragionevole sicurezza, indicano che la probabilità che si verifichino danni gravi (morti) sia dell'ordine di 100-200 per milione di persone irradiate uniformemente su tutto il corpo con la dose unitaria di 1 rem in un tempo relativamente breve. Purtroppo le stesse malattie gravissime insorgono nella specie umana soprattutto «spontaneamente» con frequenza molto elevata e cioè circa 200 000 per milione di persone in tutta la loro vita, e quindi per milioni di morti; risulta pertanto estremamente difficile individuare univocamente i casi dovuti anche a un elevato aumento artificiale della radioattività ambientale corrispondente grosso modo a una moltiplicazione per dieci dell'esposizione dovuta alla radioattività naturale (1 rem/anno contro poco più di 0,1 rem/anno).

Paradossalmente questa impostazione cautelativa del settore nucleare ha innescato nell'opinione pubblica dei paesi occidentali una controreazione negativa, nel senso che molti conoscono, e quindi temono, gli effetti della radioattività sulla salute, mentre l'opinione pubblica non conosce gli effetti tossici di molti elementi e sostanze a bassa concentrazione. Ma anche gli esperti in tossicologia hanno purtroppo poche conoscenze su tale argomento.

Per concludere la serie di riflessioni relative all'impatto ambientale delle diverse fonti energetiche è opportuno richiamare l'attenzione sulla necessità di seguire una impostazione rigorosa, anche se la scienza attuale si mostra impotente a fornire risposte esaurienti a quesiti legittimi e basilari sugli effetti sulla salute pubblica delle diverse fonti energetiche.

L'energia è uno dei supporti indispensabili al progresso dell'umanità. I flussi crescenti di energia richiesti dall'uomo pongono gravissimi problemi di approvvigionamento delle materie prime e di effetti sull'ambiente. È possibile e auspicabile tendere a una riduzione dei danni ambientali e dei rischi connessi ai processi di produzione energetica, ma è impossibile inseguire l'utopica chimera del rischio uguale a zero e dell'impatto ambientale nullo.

È stato ampiamente dimostrato che non esistono energie pulite e sicure contrapposte a energie sporche e ad alto rischio. Anche le fonti talora caldeggiate da alcuni ecologi provocano infatti gravi effetti negativi sull'ambiente e sull'uomo; quanto esposto sull'impatto ambientale dell'energia geotermica e da fusione è in proposito illuminante.

Una impostazione rigorosa dell'ampia problematica connessa agli effetti sull'ambiente delle diverse fonti energetiche è assolutamente necessaria per una responsabile pianificazione energetica. Tale impostazione rigorosa deve essere basata su quei dati di riferimento che sono oggettivi e assolutamente non obiettabili, al fine di approfondire ed estendere un dibattito sulle scelte energetiche che sia costruttivo.

In primo luogo l'esame degli effetti ambientali deve essere esteso a tutte le fasi del ciclo energetico, dall'estrazione mineraria delle materie prime necessarie, al loro trasporto e stoccaggio, alla produzione di energia, sino alla sistemazione definitiva di tutti i residui ottenuti nel processo. Per fare un esempio pertinente, nel dibattito in corso nel nostro paese si tende a trascurare il grave impatto ambientale causato dall'estrazione mineraria del carbone, per il semplice motivo che il carbone utilizzato non viene prodotto, ma importato.

La valutazione dell'impatto ambientale delle diverse fonti deve inoltre fare riferimento alle stesse potenze installate. Non ha supporto scientifico, ad esempio, il confronto proposto in assoluto tra l'impatto ambientale di una centrale geotermica da 30 MWe (del tipo di quella che verrà installata presumibilmente ai Campi Flegrei) e quello di una centrale elettronucleare da 2000 MWe o a carbone da 2500 MWe, centrali cioè con potenza quasi 100 volte più elevata.

È estremamente arduo - per non dire impossibile - effettuare confronti responsabili tra effetti che sono ben conosciuti, ed effetti possibili e altamente probabili, ma non ben noti nella loro entità. Ci si riferisce in particolare agli effetti della radioattività, da un lato, e a effetti di elementi e composti tossici non radioattivi dall'altro. All'inizio di questo paragrafo è stata presentata la problematica relativa alle relazioni tra basse dosi di radioattività ed effetti sulla salute. In questo specifico campo si dispone di punti di riferimento precisi, nel senso che si conosce con sicurezza l'ordine di grandezza dei rischi connessi ai diversi livelli di radioattività. Nel campo della tossicologia delle basse dosi di composti ed elementi tossici non radioattivi non si conosce quasi nulla circa i rischi nel medio e nel lungo termine. Il fatto che non si conosca l'entità del rischio, perché non sono state effettuate le ricerche necessarie, non riduce automaticamente a zero l'entità del rischio stesso. Recentissime ricerche sulle conseguenze sulla salute pubblica delle iniziative previste nell'ambito del risparmio energetico per migliorare l'isolamento delle abitazioni forniscono un esempio significativo al riguardo.

Un accurato studio effettuato da D. Dickson dell'Università della California indica che l'aumento del tenore in ^{222}Rn nelle abitazioni causato dal ridotto ricambio di aria con l'esterno per l'isolamento delle nuove costruzioni provocherà un aumento dei decessi per cancro polmonare da 20 a 200 per anno per ogni milione di abitanti. Uno studio dell'Environmental Protection Agency ancora più recente rincara la dose, affermando che i dati disponibili indicano che il 10 per cento dei cancri polmonari negli Stati Uniti sono provocati dal rado all'interno delle abitazioni e che il programma di aumentare l'isolamento delle abitazioni condurrà a un aumento di 80-100 casi di morte per anno per milione di persone.

I risultati di queste ricerche sono invero emblematici di una situazione ben più generale. Il fatto che siano state sviluppate le tecnologie per ottenere misure affidabili anche ai bassissimi livelli di rado e che si conosca grosso modo per tale elemento la relazione tra tenore nell'aria ed effetti sulla salute dell'uomo permette di fornire valutazioni attendibili sugli effetti sulla salute pubblica di una operazione molto semplice quale l'aumento dell'isolamento delle abitazioni. Per altri elementi e composti tossici non radioattivi, che pure tendono ad arricchirsi in un ambiente abitato ben isolato, non si conoscono i tassi di arricchimento e neppure la relazione tra dosi ed effetti sulla salute, e il rischio connesso viene quindi di fatto tranquillamente ignorato come se fosse uguale a zero. E tra questi ultimi composti ce ne sono alcuni altamente tossici, che per più tendono ad accumularsi nell'organismo umano e a danneggiare suoi organi importanti; è questo ad esempio il caso dei vapori di mercurio che si originano dai termometri che vengono rotti accidentalmente nelle abitazioni.

Altro aspetto, che merita una attenzione ben maggiore di quella che normalmente gli viene attribuita, è costituito dalla sostanziale unitarietà e globalità dell'ambiente, con le profonde interazioni esistenti tra e all'interno delle diverse sfere geochimiche, e con flussi elevati e veloci di sostanze nelle sfere più mobili (atmosfera e idrosfera) e tra queste e le altre.

È stato dimostrato che l'uomo ha già provocato variazioni ambientali alla scala dell'intero nostro globo, sia per quel che riguarda la composizione dell'atmosfera sia per quel che riguarda l'inquinamento da metalli tossici e da sostanze organiche tossiche. Finisce quindi per essere una illusione senza prospettive occuparsi solo del proprio campanile, della propria vallata o soltanto di alcuni tossici, perché la contaminazione ambientale non rispetta i confini di comuni, regioni e nazioni e si è già verificata in maniera apprezzabile per un grandissimo numero di elementi e sostanze tossiche.

Altri esempi possono chiarire meglio la gravità e la globalità della contaminazione ambientale. La acidità delle piogge che cadono sul Canada riceve un contributo determinante (circa il 50 per cento) dall'attività industriale degli Stati Uniti. E le piogge acide possono provocare gravi danni all'attività agricola canadese, che rappresenta una fonte di sostentamento importante per l'umanità intera.

Sempre nell'ambito dei problemi sollevati dalle piogge acide i preziosi monumenti di Atene e di Roma hanno subìto danni molto gravi e irreversibili negli ultimi decenni. Uno dei problemi importanti in questi casi è stabilire in che misura e fino a che punto i monumenti di Atene appartengono agli ateniesi o ai greci o non facciano piuttosto parte di un patrimonio unico di tutta l'umanità.

Per quel che riguarda la radioattività ambientale la tabella di pagina 102 riporta in sintesi il valore medio delle diverse fonti. A questo proposito può essere un non senso il cercare di ridurre il lieve incremento di radioattività ambientale derivante dall'installazione di centrali nucleari nel territorio italiano, quando l'installazione di centrali a carbone può portare a un incremento della radioattività ambientale dello stesso ordine, nel mentre la Francia ha in via di completamento la supercentrale al plutonio Superphenix non molto lontano dal confine italiano, e la Russia persegue un programma di forte sviluppo di una filiera di reattori del tutto simile a quella francese.

Le scelte vanno quindi effettuate secondo l'ottica della salvaguardia sia dei diritti locali come pure dei diritti delle comunità nazionali e sovrannazionali. Tutta l'umanità si trova a navigare sulla stessa barca e i problemi ambientali che si trova - e sempre più si troverà nel futuro - ad affrontare non possono essere risolti nascondendo o volutamente sottostimando la gravità dei problemi stessi. È necessario prendere coscienza delle rigide condizioni al contorno, che sussistono attualmente, e che presumibilmente diventeranno ben più restrittive nel prossimo futuro per l'aggravarsi dei problemi relativi all'approvvigionamento energetico e del relativo stato di salute dell'ambiente.

Sono già state osservate le modificazioni ambientali e la serie di effetti negativi sull'ambiente e sulla salute pubblica che ne derivano. L'imperativo categorico che deve essere rispettato è quello di non lasciare in eredità alle generazioni future un ambiente irreversibilmente alterato, con la possibilità di scatenare dei processi di reazioni a catena, quali quelle relative a cambiamenti consistenti del clima, che possono rendere in un futuro non lontanissimo questo nostro pianeta deteriorato a tal punto da essere inabitabile.

La produzione di energia è una delle maggiori responsabili delle alterazioni ambientali. Anche se si deve cercare di ridurre la gravità della contaminazione ambientale mediante il risparmio, mediante l'applicazione delle migliori tecnologie disponibili per la riduzione delle emissioni di sostanze tossiche verso l'ambiente e mediante l'approfondimento delle ricerche e l'estensione dei controlli sia nel settore tossicologico sia in quello ambientale, bisogna anche accettare gli effetti negativi prodotti dall'energia, questa utile ma scomoda compagna di viaggio dell'umanità. I problemi da affrontare e da risolvere sono comunque di una tale gravità e vastità che richiedono l'impegno e l'accordo, non solo a livello nazionale, ma anche, e per molti aspetti soprattutto, a livello sovrannazionale e globale.

In fondo il problema della produzione energetica, e industriale in generale, non è dissimile dal problema costituito dalle fogne. Non si dispone della bacchetta magica per risolvere il problema dello smaltimento dei rifiuti solidi e liquidi. Sappiamo solo con sicurezza che il problema non può essere risolto né vietando la produzione dei rifiuti, né tanto meno scaricando le fogne nel territorio dei comuni o delle nazioni confinanti.

La percezione anortoscopica

Una figura mostrata una parte alla volta attraverso una stretta fenditura può venir percepita nella sua interezza. Questo effetto dipende dalla «pittura» dell'immagine sulla retina o da eventi psicologici più profondi?

di Irvin Rock

Da molto tempo la psicologia sperimentale studia il modo in cui si percepisce visivamente la forma degli oggetti, ma non è ancora riuscita ad appurarlo perfettamente. Per certi versi l'occhio è come una macchina fotografica, in cui un oggetto di una certa forma dà origine a una immagine della stessa forma sulla superficie della retina. La semplice analogia, tuttavia, viene meno in situazioni in cui la presenza di tale immagine non è né sufficiente né necessaria per dar conto della forma percepita. Allo scopo di saperne di più sui processi percettivi i ricercatori hanno approntato, nel corso degli anni, una varietà di tecniche di laboratorio. Io e i miei colleghi dell'Institute for Cognitive Studies della Rutgers University ci siamo recentemente concentrati su un metodo che richiede la visione di una figura in sezioni successive attraverso una stretta fenditura. Consapevoli dell'insolita natura di questa situazione sperimentale abbiamo adottato il termine ottocentesco *anortoscopica* («vista in maniera anormale») per caratterizzare la percezione in questi casi.

Uno dei primi problemi che si presentano con l'analogia occhio-macchina fotografica è che essa non prende in considerazione la questione di come le componenti dell'immagine retinica vengono organizzate. Consideriamo la disposizione di tre punti non allineati. Essi saranno visti come i vertici di un triangolo soltanto se sono raggruppati in una unità dal sistema percettivo; in presenza di altri punti la loro «triangolarità» può essere meno evidente. Anche una figura ben conosciuta composta di linee può non essere riconosciuta se il sistema percettivo associa le linee con altre linee per formare differenti raggruppamenti. Inoltre vi è la questione dell'organizzazione in figura e sfondo: una forma familiare non può essere percepita - anche se una sua immagine è presente sulla retina - se qualche altra parte del campo visivo è considerata come se fosse una figura.

Un altro problema è quello dell'orientamento. Se si percepisce scorrettamente la localizzazione dell'alto, del basso e dei lati di una figura, la sola presenza della sua immagine sulla retina non ci assicura che la figura sarà riconosciuta. Anche l'orientamento di una figura nella terza dimensione può costituire un fattore della percezione della forma. Per esempio, un cerchio, visto obliquamente, proietta un'immagine ellittica sulla retina; ma può nondimeno apparire circolare in condizioni appropriate.

Tutti questi esempi indicano che l'esistenza di una precisa immagine retinica non è una spiegazione sufficiente della percezione di una forma. Si devono invocare altri fattori. Anzi, a volte una precisa immagine retinica non è neppure un fattore necessario. Si può mostrare facilmente che anche se soltanto alcune parti di una figura sono presenti, si possono nondimeno percepire anche le altre, come se anche queste fossero presenti. Si consideri un semplice disegno al tratto di due rettangoli, uno dei quali appare sovrapporsi all'altro e nasconderlo; la figura può essere vista come formata da due rettangoli completi anche se l'immagine di uno è incompleta. In altri casi si vedono dei contorni (chiamati contorni soggettivi o illusori), nonostante l'assenza nell'immagine retinica della brusca transizione dalla luce al buio che si considera di solito la condizione stimolo per percepire un contorno (si veda l'articolo *Contorni soggettivi* di Gaetano Kanizsa, in «Le Scienze», n. 96, agosto 1976).

Negli ultimi due esempi, almeno qualcuno dei contorni della figura è presente nell'immagine retinica. In altri esperimenti è possibile percepire una forma senza alcun contorno nell'immagine retinica. Con l'aiuto di uno stereoscopio si può presentare a ciascun occhio separatamente una distribuzione casuale di elementi grafici in modo che nessuna figura sia visibile con uno solo dei due occhi. Tuttavia se nell'immagine presentata a un occhio un gruppo di elementi è spostato in senso orizzontale rispetto allo stesso gruppo nell'immagine presentata all'altro occhio, in questo caso il sistema percettivo può rilevare la somiglianza delle due configurazioni e inferire che c'è una forma o davanti al resto dell'immagine presentata o dietro. Qui la forma percepita non emerge dai contorni dell'immagine retinica presente in ciascun occhio, ma dai confini di una regione il sistema percettivo inferisce l'esistenza di un'entità nella scena (si veda l'articolo *Percezione visiva delle strutture* di Bela Julesz, in «Le Scienze» n. 84, agosto 1975).

Un altro esempio ha un legame più stretto con le mie ricerche attuali. Supponiamo che un punto luminoso percorra una determinata traiettoria in una camera oscura e che si addestri un osservatore a seguirlo con gli occhi. L'immagine del punto cadrebbe sempre all'incirca nello stesso luogo della retina (specificatamente la fovea); ciononostante l'osservatore percepirebbe la forma della traiettoria. Questa impressione sarebbe molto vicina alla percezione di una forma e, tuttavia, sulla retina non vi sarebbe alcuna immagine estesa corrispondente alla traiettoria del punto. Quando Fred Halper e io abbiamo fatto questo semplice esperimento nel nostro laboratorio alla Rutgers University, abbiamo trovato che gli osservatori percepiscono la forma di questa traiettoria in maniera altrettanto buona di quando tengono fermi gli occhi così che l'immagine del punto si muova lungo la retina.

Si potrebbe tuttavia sostenere che la percezione di una traiettoria non è la stessa cosa della percezione di una forma. Dopo tutto, nell'ultimo caso si vede un contorno, non soltanto un punto che si muove. Tuttavia, con una piccola modifica alla procedura sperimentale, c'è un modo per ottenere un'efficace impressione di forma. Invece di inseguire un punto che si muove in una camera oscura il soggetto osserva una stretta fenditura su una superficie opaca; la fenditura si muove lungo una linea che disegna una figura *(si veda l'illustrazione nella pagina a fronte)*. Se l'osservatore segue attentamente la fenditura, questa stimola soltanto una porzione ristretta della retina e di conseguenza l'immagine della porzione visibile della figura si muove su e giù all'interno di questa porzione di retina. Per converso la figura può essere mossa avanti e indietro, dietro una fenditura che rimane immobile; se gli occhi dell'osservatore in questo caso rimangono fermi, fissando la fenditura, ancora una volta la fenditura stessa stimola solo una ristretta area della retina. In nessun momento, in entrambi i casi, sulla retina c'è un'immagine che rassomigli alla forma della figura. Ciononondimeno abbiamo scoperto che, date certe condizioni, le figure sono facilmente percepite.

Nel secolo scorso parecchi ricercatori, tra i quali Joseph Plateau, Fritz Zöllner e Hermann von Helmholtz, hanno impiegato un dispositivo simile per studiare determinati fenomeni visivi. Essi parlavano della loro procedura in termini di presentazione anortoscopica, presumibilmente in quanto si trattava di un metodo inusuale di stimolazione visiva. Il dispositivo sperimentale, chiamato anortoscopio, consisteva di due dischi montati assialmente uno dietro l'altro. I dischi potevano ruotare a differente velocità in direzioni opposte. L'osservatore attraverso fenditure sul disco anteriore vedeva delle figure sul disco posteriore, percependo così determinate distorsioni illusorie delle figure stesse.

La presentazione anortoscopica è un metodo inusuale di stimolazione visiva, adottato dall'autore e dai suoi colleghi per verificare l'ipotesi della «pittura» retinica, la quale sostiene che sulla retina l'immagine si forma per parti successive, come un quadro sotto il pennello del pittore. Con questa tecnica, l'osservatore vede sezioni successive di una figura attraverso una stretta fenditura. Vi sono due condizioni sperimentali possibili: o la fenditura si muove davanti a una figura immobile *(a sinistra)* oppure una figura si muove dietro una fenditura immobile *(a destra)*.

Queste vecchie ricerche sono state dimenticate fino al 1965, quando sono state riscoperte da T. E. Parks dell'Università della California a Davis. Sia nell'Ottocento sia dopo la riscoperta sono nate controversie su come vada interpretato l'effetto anortoscopico. Helmholtz e altri pensavano che sulla retina fosse sempre impressa un'immagine estesa della figura che si trovava dietro la fenditura. Questa interpretazione da allora è nota con il nome di ipotesi della «pittura retinica», poiché si basa sull'idea che un'immagine si formi sulla retina gradualmente, una porzione alla volta. Quando la fenditura si muove davanti a una figura immobile, secondo questa ipotesi l'osservatore dovrebbe tenere gli occhi fermi e non seguire la fenditura; quando invece la figura si muove dietro una fenditura immobile l'osservatore dovrebbe spostare gli occhi avanti e indietro. Helmholtz sosteneva che è possibile che l'osservatore non sempre sappia se i suoi occhi sono fermi o in movimento.

Se l'ipotesi della «pittura retinica» è corretta, la percezione della forma sotto siffatte condizioni difficilmente potrebbe sorprendere. In entrambi i casi citati, allora sarebbe presente un'immagine retinica estesa di una figura, e la sola differenza tra la percezione anortoscopica e quella normale consisterebbe nel fatto che nella prima l'immagine si formerebbe per segmenti presentati in successione anziché simultaneamente. Inoltre, poiché le cellule sensoriali della retina tendono a prolungare la scarica per breve tempo dopo che la stimolazione è terminata (fenomeno conosciuto come persistenza neurale), l'immagine presentata in segmenti successivi non differirebbe in maniera significativa da un'immagine che si forma tutta in una volta, purché le presentazioni si susseguano abbastanza velocemente. Ci si dovrebbe quindi aspettare che l'osservatore percepisca l'intera immagine anortoscopica come se si fosse stabilita normalmente.

Per valutare l'ipotesi della «pittura retinica» bisogna sapere esattamente che cosa fanno gli occhi dell'osservatore durante una presentazione anortoscopica. Pertanto nei nostri esperimenti abitualmente registriamo con un videoregistratore gli occhi dell'osservatore durante la presentazione. In un esperimento condotto da Joseph Di Vita e dall'autore, gli osservatori guardavano figure costituite da linee curve lunghe circa 14 centimetri, attraverso una fenditura larga 0,31 centimetri, dalla distanza di 76 centimetri. In un caso la figura si muoveva avanti e indietro per la durata di 5 cicli, nell'altro era la fenditura a muoversi. La velocità della figura o della fenditura era di mezzo ciclo al secondo, inferiore a quella adottata da altri che hanno indagato sulla percezione anortoscopica. Le registrazioni indicavano in ogni momento la direzione dello sguardo in riferimento a ciò che veniva mostrato. Si chiedeva all'osservatore di premere il pulsante soltanto nei momenti in cui percepiva una figura dietro alla fenditura. Il segnale veniva registrato su nastro insieme con la registrazione dei movimenti oculari.

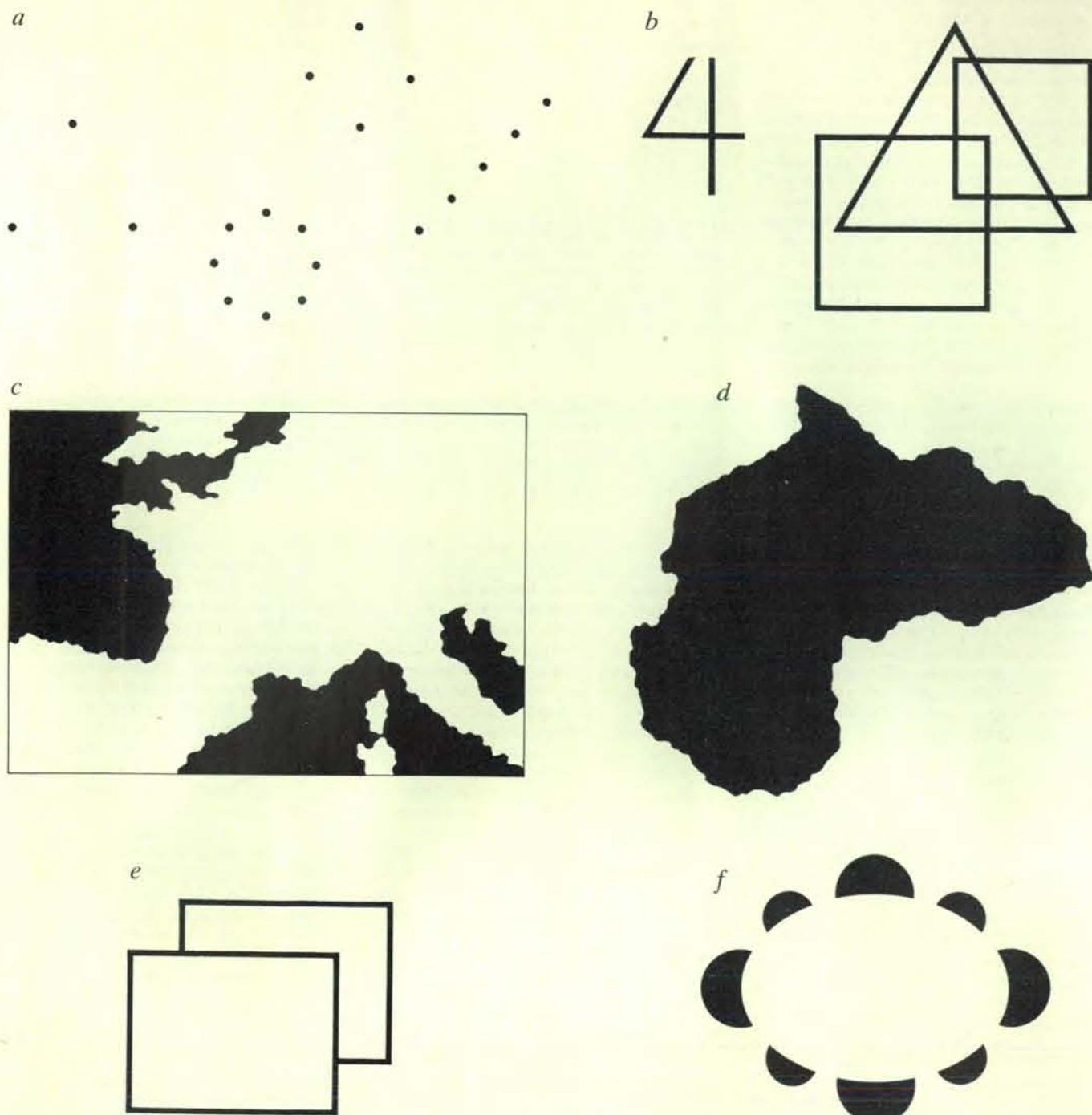
I dati hanno mostrato che, quando la fenditura era ferma, la maggior parte delle volte gli osservatori hanno tenuto fermi gli occhi, sia che fosse stato detto loro di fare così o meno, sia che fosse stato dato loro un punto da fissare sulla fenditura o meno. Ovviamente gli occhi non rimanevano perfettamente fermi, ma gli occasionali movimenti da una parte e dall'altra (conosciuti come movimenti saccadici) non erano sufficienti a rendere conto della percezione della forma. Gli osservatori non sempre hanno percepito una figura, ma molte volte sì, e quando i loro occhi erano più o meno fermi.

Questi risultati sono stati recentemente confermati da Robert Fendrich e Arien Mack della New School for Social Research. Nei loro esperimenti l'immagine presentata era stabilizzata in maniera tale che, anche se gli occhi si muovevano, l'immagine della fenditura non poteva dislocarsi lungo la retina. Ciononondimeno anche i loro osservatori percepivano una figura in movimento dietro la fenditura.

Quando nei nostri esperimenti la fenditura si muoveva davanti a una figura immobile, tipicamente i soggetti muovevano gli occhi insieme con la fenditura. Tuttavia di solito si verificava un ritardo: la fenditura sopravanzava rapidamente gli occhi e rimaneva in testa fino a che non raggiungeva la sua posizione finale, dopo di che veniva rapidamente raggiunta dagli occhi. Era chiaro ciononondimeno, che l'immagine della fenditura non si muoveva abbastanza lontano sulla retina così da spiegare la percezione della forma; anzi, l'immagine retinica non si muoveva neppure coerentemente nella giusta direzione. In questa situazione era quasi sempre percepita una figura, e la sua dimensione era percepita più o meno correttamente.

Altri esperimenti, condotti nel nostro laboratorio e altrove, indicano che l'effetto anortoscopico potrebbe risultare dalla pittura retinica solamente se l'immagine della fenditura si spostasse sulla retina molto velocemente. L'impressione di una figura completa potrebbe essere spiegata da una persistenza neurale solamente con una velocità prossima a 5 percorsi completi della retina al secondo (cinque volte la velocità usuale). D'altra parte, nel caso in cui sia la figura a muoversi, la pittura retinica dovrebbe essere

basata sul movimento degli occhi, che pertanto dovrebbero spostarsi avanti e indietro a un ritmo rapido. Ma gli osservatori non sono capaci di muovere gli occhi così velocemente.

Sulla base di questa e di altre prove di fatto dobbiamo concludere che la pittura retinica non fornisce una spiegazione sufficientemente generale della percezione anortoscopica. Quale è allora la base della percezione della forma negli esperimenti di questo tipo? Consideriamo ancora l'inseguimento di un punto luminoso in una camera oscura. Sebbene l'immagine del punto non si muova sulla retina, il sistema percettivo riceve informazioni riguardo alla traiettoria del punto poiché gli occhi devono muoversi per inseguirlo. Il sistema percettivo, nel giudicare dallo stimolo retinico la posizione del punto nello spazio, tiene conto della posizione degli occhi. Così, la traiettoria percepita del punto viene derivata dall'insieme delle posizioni percepite del punto stesso mentre si muove.

Il ruolo limitato dell'immagine retinica nella percezione visiva della forma è illustrato attraverso sei esempi. In *a* i tre punti a sinistra sono visti come i vertici di un triangolo soltanto se il sistema percettivo li organizza in una unità; in presenza di altri punti *(a destra)* il triangolo può non essere percepito. In *b* il numero quattro *(a sinistra)* non sarà visto come tale se il sistema percettivo associa le linee che lo costituiscono con altre linee in modo tale da formare differenti raggruppamenti *(a destra)*. In *c* la forma familiare dell'Europa occidentale non verrà percepita se lo sfondo (in questo caso i mari che la circondano) è considerato come figura. Parimenti in *d* il continente africano può non essere riconosciuto se non si è consapevoli che la parte superiore della figura (cioè il nord) si trova a sinistra. Questi quattro esempi mostrano come la semplice presenza di una specifica immagine retinica non renda conto a sufficienza di ciò che viene percepito. In *e* due forme rettangolari, di cui una si sovrappone all'altra e la occlude, possono essere viste come due rettangoli completi, uno dei quali nasconde in parte l'altro, anche se uno di essi è di fatto incompleto. In *f* un contorno soggettivo (o illusorio) è percepito facilmente anche in assenza della brusca transizione dal chiaro allo scuro che generalmente si considera la condizione stimolo perché si abbia la percezione dei contorni. Gli ultimi due esempi servono a rendere evidente come non sia necessaria un'immagine retinica precisa o completa per spiegare che cosa viene percepito.

Generalizzando questa argomentazione alla percezione di tutte le forme, si può affermare che quando un'immagine estesa di una figura è presente sulla retina l'osservatore percepisce ogni punto della figura come se fosse in un determinato luogo. Per cui si potrebbe supporre che la forma percepita è il risultato di una sintesi mentale delle posizioni percepite di tutti i punti che costituiscono la forma. Tale processo potrebbe essere organizzato gerarchicamente, con punti sintetizzati in linea e la disposizione delle linee le une rispetto alle altre sintetizzata in figure. Se così fosse, potrebbe essere l'insieme delle posizioni percepite ad acquisire rilevanza per l'immagine retinica di una figura. In altre parole, forse, la funzione dei contorni retinici di una figura è quella di stabilire un insieme di posizioni che congiuntamente costituiscano una figura.

La stessa argomentazione si può applicare alla percezione anortoscopica. Il sistema percettivo integra le informazioni sulle posizioni delle parti della figura, così come esse appaiono successivamente attraverso la fenditura. Quando la fenditura si muove su una figura immobile, l'elemento visibile a una estremità della figura è caratterizzato, diciamo, come «all'estremità inferiore della fenditura e alla sinistra», mentre un attimo dopo è caratterizzato come «più in alto nella fenditura e diritto davanti». L'integrazione delle direzioni conduce alla percezione di una linea obliqua che si inclina verso l'alto a destra. Perciò si può dire che, sia la percezione anortoscopica delle forme sia quella normale, implicano un'integrazione di tutte le posizioni date delle parti di una figura, le une rispetto alle altre, in una descrizione del tutto.

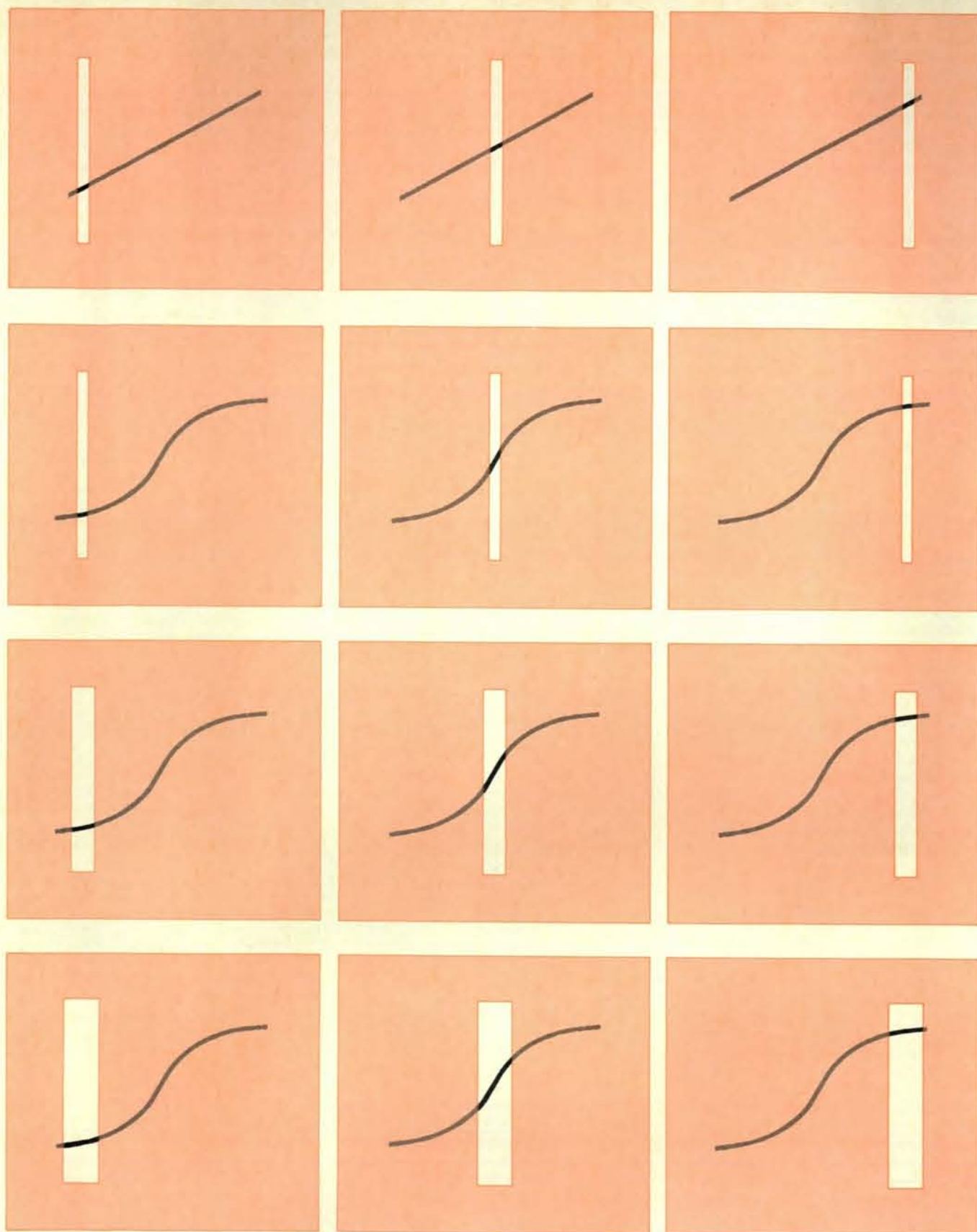
L'autore e uno dei suoi colleghi, Alan L. Gilchrist, hanno messo a punto una tecnica di simulazione per studiare il fenomeno della percezione anortoscopica. In questo caso la tecnica è stata usata per verificare il fatto che, perché sia visibile una figura nella situazione anortoscopica, la parte visibile della figura deve estendersi per tutta la larghezza della fenditura. Un piccolo segmento di linea, disegnato su di un pezzo di plastica trasparente, è attaccato mediante sottili bacchette a un sistema di rotelle che scorrono su di una camma non visibile attraverso la fenditura. La parte visibile del segmento di linea si muove su e giù e contemporaneamente cambia la sua inclinazione proprio come farebbe la parte visibile di una linea complessa che scorre dietro. Quando la fenditura è più stretta della lunghezza del segmento spesso si può percepire una figura *(a sinistra)*. Tuttavia quando la fenditura è più larga del segmento non si percepisce mai la figura *(a destra)*.

Questa spiegazione risulta incompleta per diversi motivi. In primo luogo, la direzione in relazione all'osservatore è data direttamente nella condizione in cui è la fenditura a muoversi, ma non nella condizione in cui si muove la figura. Nel secondo caso la fenditura rimane ferma diritta davanti all'osservatore. Donde provengono allora le informazioni circa le diverse posizioni occupate dalle parti della figura che si mostrano l'una dopo l'altra? Evidentemente si deve prima ipotizzare che qualcosa si sta muovendo dietro la fenditura. Una volta che si è accettata questa idea, la successione temporale degli stimoli può essere trasformata in una configurazione spaziale. In seguito, come nella condizione della fenditura che si muove, il sistema percettivo può integrare la successione di stimoli in una forma basata sull'insieme di direzioni delle parti, le une rispetto alle altre.

La nostra analisi suggerisce un'altra ragione della non esaustività dell'ipotesi della sintesi delle posizioni. Spesso degli osservatori, nella presentazione anortoscopica, non percepiscono una figura. Tale risultato può in qualche caso essere attribuito al genere di figura che noi adoperiamo, poiché altri ricercatori non hanno riferito di insuccessi di questo tipo. Noi abbiamo scelto deliberatamente figure con numerose caratteristiche distintive: sono linee singole e sono formate in modo tale che solo un elemento alla volta sia visibile nella fenditura; sono curve continue e regolari piuttosto che configurazioni con improvvisi cambiamenti di direzione; sono linee continue piuttosto che punteggiate o tratteggiate; non rappresentano oggetti familiari. Altri hanno adoperato figure come quadrati, cerchi ed ellissi, o disegni al tratto di oggetti familiari come, ad esempio, animali.

Noi pensiamo che il nostro genere di figura sia ambiguo per il fatto che la sua presentazione anortoscopica potrebbe logicamente rappresentare sia un elemento che si muove verticalmente in una fenditura, sia una figura estesa che viene rivelata attraverso la fenditura stessa. Perciò ci siamo posti il problema di analizzare quali fattori siano importanti nella percezione di una figura, dal momento che non sempre viene percepita una figura. La nostra ipotesi guida è stata che la percezione di una figura in condizioni anortoscopiche sia basata su un processo inconscio di soluzione di problemi. Ancor prima che abbia inizio una qualsiasi attività di sintesi delle posizioni, il sistema percettivo deve ipotizzare che dietro la fenditura sta una figura estesa, e che questa viene rivelata in sezioni successive.

Per chiarire questa ipotesi consideriamo ciò che si vedrebbe se la figura fosse luminosa e osservata in una camera oscura, così che la fenditura non sia visibile. Nel caso della fenditura immobile tutto quello che un osservatore vedrebbe sarebbe un punto di luce che si muove su e giù. Nella condizione di fenditura mobile l'osservatore vedrebbe un punto di luce che si muove percorrendo una traiettoria. In nessuno dei due casi si potrebbe descrivere il percetto (l'entità percepita) come una figura estesa. Perciò sembra che l'osservatore debba accorgersi che ciò che vede in un dato momento lo vede attraverso una stretta fenditura e che ciò che resta della figura è nascosto da una superficie opaca. Diversamente dalla soluzione di problemi come atto di pensiero, la soluzione di problemi nella percezione richiede un sostegno da parte dello stimolo. Per percepire una figura anortoscopicamente non è sufficiente sapere che si sta guardando attraverso una fenditura; la fenditura deve essere vista.

In certe condizioni l'osservatore può anche «inventare» una fenditura quando non è affatto visibile. Fendrich e Mack hanno trovato che, se un cerchio viene mosso dietro una fenditura invisibile e viene fornito il suggerimento opportuno, degli osservatori, che all'inizio vedono le parti visibili del cerchio muoversi verticalmente, successivamente percepiscono una figura che si muove orizzontalmente dietro una fenditura illusoria. Il loro esperimento fornisce un buon esempio di come il sistema percettivo tende a razio-

Quattro sequenze di presentazione (ciascuna sequenza presenta tre fasi successive) dimostrano la forte tendenza a percepire una linea mostrata anortoscopicamente come se fosse un segmento sottoposto a uno spostamento verticale piuttosto che una linea continua che si rivela orizzontalmente. Quando una linea retta obliqua viene mostrata anortoscopicamente *(in alto)*, essa non sembra mai una figura estesa in movimento probabilmente perché non c'è alcun indizio di un qualche cambiamento nell'identità del segmento. Quando nelle medesime condizioni *(seconda dall'alto)* viene presentata la figura di una linea curva, il segmento visibile accelera e decelera muovendosi verticalmente entro la fenditura, ma questa ulteriore informazione non sembra modificare la percezione. Anche quando si raddoppia la larghezza della fenditura, così che si possono riconoscere l'inclinazione e la variazione di inclinazione del segmento visibile *(terza dall'alto)*, la maggior parte degli osservatori non percepisce ancora una figura. Solo quando la figura è allargata ulteriormente così che diventano ravvisabili la curvatura e le variazioni di curvatura *(in basso)*, viene percepita una figura. Tutte le figure sono state viste da una distanza di circa un metro.

nalizzare una soluzione, ricavata dall'attitudine alla soluzione di problemi, qualora lo stimolo non fornisca un sostegno sufficiente.

Non solo è necessario che la fenditura sia visibile: la parte visibile della fenditura deve pure estendersi completamente da una parte all'altra della fenditura stessa. Per chiarire questo punto Alan L. Gilchrist e io abbiamo escogitato una tecnica per simulare la presentazione anortoscopica. Invece di una figura, dietro la fenditura c'era solamente un piccolo segmento di linea disegnato su una plastica trasparente. La plastica era attaccata mediante sottili bacchette a un insieme di rotelle che scorrevano su di una camma non visibile attraverso la fenditura; la camma aveva la forma della figura che doveva essere simulata *(si veda l'illustrazione a pagina 113)*.

Date certe condizioni una simulazione con tale dispositivo conduce alla percezione di una figura. In questa situazione, tuttavia, abbiamo ampliato la fenditura in modo che il segmento di linea visibile non la occupasse tutta, ma piuttosto rimanesse isolato al suo interno. Sebbene il segmento si muovesse verticalmente e cambiasse la sua inclinazione, proprio come avveniva quando la fenditura era più stretta, non veniva mai percepita una figura. Lo stimolo evidentemente violava la logica della soluzione: cioè che l'osservatore stava vedendo una figura continua che veniva rivelata attraverso la fenditura.

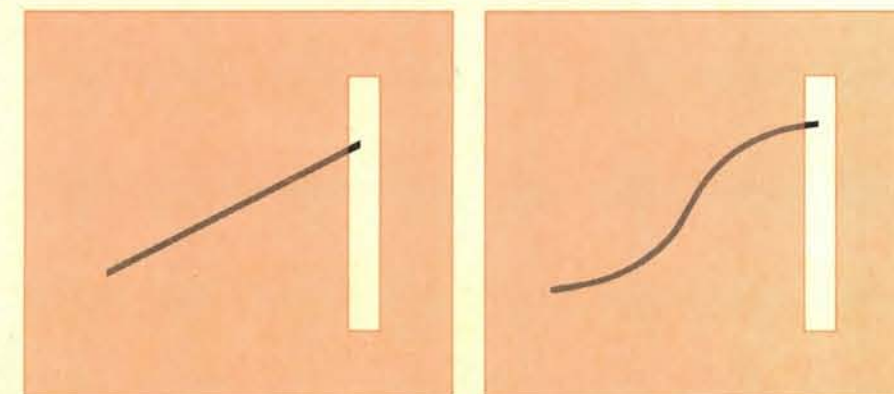
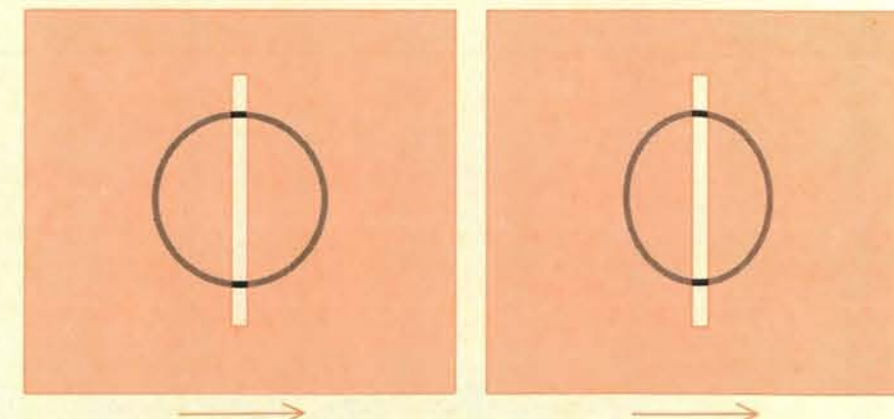
Consideriamo, adesso, il genere di figura presentato dietro la fenditura immobile. Se la figura è una linea retta orientata obliquamente e le estremità non passano attraverso la fenditura (che è larga circa 0,15 centimetri ed è vista da una distanza di circa 1 metro), lo stimolo consiste in un segmento di linea che è sottoposto a uno spostamento verticale. Tale stimolo potrebbe derivare da uno di questi due eventi fisici: un segmento che si muove su e giù a velocità uniforme o una linea obliqua che si muove avanti e indietro orizzontalmente a una velocità uniforme.

In queste condizioni di ambiguità Hans Wallach dello Swarthmore College ha dimostrato tempo fa che la forma dell'apertura ha una funzione importante nel determinare ciò che viene percepito. Vi è una forte tendenza a percepire la linea come se si muovesse parallelamente all'asse maggiore dell'apertura, probabilmente perché con una fenditura stretta non esiste alcuna informazione che suggerisca un qualsiasi cambiamento nell'identità del segmento. Corrispondentemente, abbiamo trovato che sotto tali condizioni non viene mai percepita una figura estesa e in movimento.

Supponiamo che la figura sia una linea curva ma che le condizioni di presentazione siano per il resto identiche. Qui il solo fattore nuovo è che il segmento visibile accelera e decelera mentre si muove verticalmente. Ann Corrigan e io abbiamo trovato che tale informazione non modifica il risultato. Tutti gli osservatori percepiscono un segmento che si muove verticalmente e non una figura.

Supponiamo ancora che la larghezza della fenditura venga portata a 0,31 centimetri. In tali condizioni abbiamo trovato che l'inclinazione del segmento visibile e il cambiamento dell'inclinazione possono essere riconosciuti attraverso la fenditura. Comunque anche con tale informazione e la correlata accelerazione e decelerazione verticale del segmento, la maggior parte degli osservatori non percepisce ancora una figura. A questo punto si potrebbe concludere che la tendenza, notata da Wallach, di un segmento a mantenere la sua identità è così forte che si presenta anche nel caso che il segmento appaia piegarsi avanti e indietro mentre sale e scende.

Solamente quando la fenditura viene ulteriormente allargata, così che la curvatura e il cambiamento nella curvatura diventano distinguibili attraverso la fenditura, l'equilibrio si sposta in favore della percezione di una figura. Perché le cose vadano così non è ancora chiaro, sebbene vengano in mente molte possibilità. Si potrebbe ipotizzare che l'accelerazione di un segmento di linea ascendente, proprio perché la sua inclinazione si sposta verso la verticale e proprio perché la sua curvatura cambia appropriatamente, rappresenterebbe una eccessiva coincidenza se non ci fosse una linea curva estesa dietro la fenditura. Una soluzione che faccia riferimento a una figura che si muove spiega perché tali caratteristiche varino insieme a questo modo, e il sistema percettivo tende a cercare tali soluzioni e a rigettare quelle che postulano variazioni per pura coincidenza simultanea.

Alternativamente, si potrebbe dire che il concetto di un segmento di linea che si muove verticalmente mentre simultaneamente si piega e cambia la sua curvatura è una percezione molto complessa. Una figura che si muove è una percezione più semplice, ed è stato argomentato che il sistema percettivo preferisca la descrizione più semplice conforme allo stimolo. Inoltre si potrebbe considerare che la combinazione di cambiamenti, inclusi i cambiamenti di curvatura, è riconoscibile

La comparsa degli estremi di una linea nella fenditura di una presentazione anortoscopica, può aiutare a percepire una forma, ma non sembra da sola sufficiente a spostare l'equilibrio in favore della percezione di una figura. Una linea retta obliqua tende ancora a essere vista come un segmento verticale anche quando sono visibili i suoi estremi *(a sinistra)*. Invece, una linea curva di cui siano visibili gli estremi è vista più facilmente come una figura *(a destra)*.

Un effetto di compressione si può frequentemente osservare nelle presentazioni anortoscopiche in cui la figura si muove dietro una fenditura immobile. La figura percepita - nel caso rappresentato in figura, un cerchio *(a sinistra)* - viene tipicamente compressa lungo il suo asse di movimento, dando origine a quella che sembra essere un'ellisse allungata in senso verticale *(a destra)*.

dal sistema percettivo, sulla base dell'esperienza passata, come la combinazione prodotta da un oggetto che passa dietro una stretta apertura.

Vi sono altri tipi di informazione, nella presentazione, che possono influenzare il risultato. Se si lascia che le estremità della figura siano viste attraverso la fenditura si può arrivare alla percezione di una forma. Chiaramente, quando le estremità sono visibili c'è un momento di informazione non ambigua, nel senso che qualcosa si sta muovendo attraverso la fenditura e non lungo la stessa. Si potrebbe ritenere tale informazione come un indizio che produce un'ipotesi di figura, ma in se stessa tale informazione non sembra essere adeguata; una linea retta obliqua tende ancora a essere vista come un segmento che si muove verticalmente, salvo nel breve momento in cui sono visibili gli estremi. Tuttavia, quando tale informazione relativa all'estremo della linea è abbinata a quella del cambiamento di inclinazione, l'equilibrio è spesso spostato verso una percezione di figura. Forse la ragione è che la percezione di un cambiamento di inclinazione tende a sostenere l'ipotesi ricavata dalla visione delle estremità della linea. L'informazione non ambigua fornita dalle estremità può essere presente in ogni parte di una figura, se presenta discontinuità nette o cambiamenti nella direzione dei contorni.

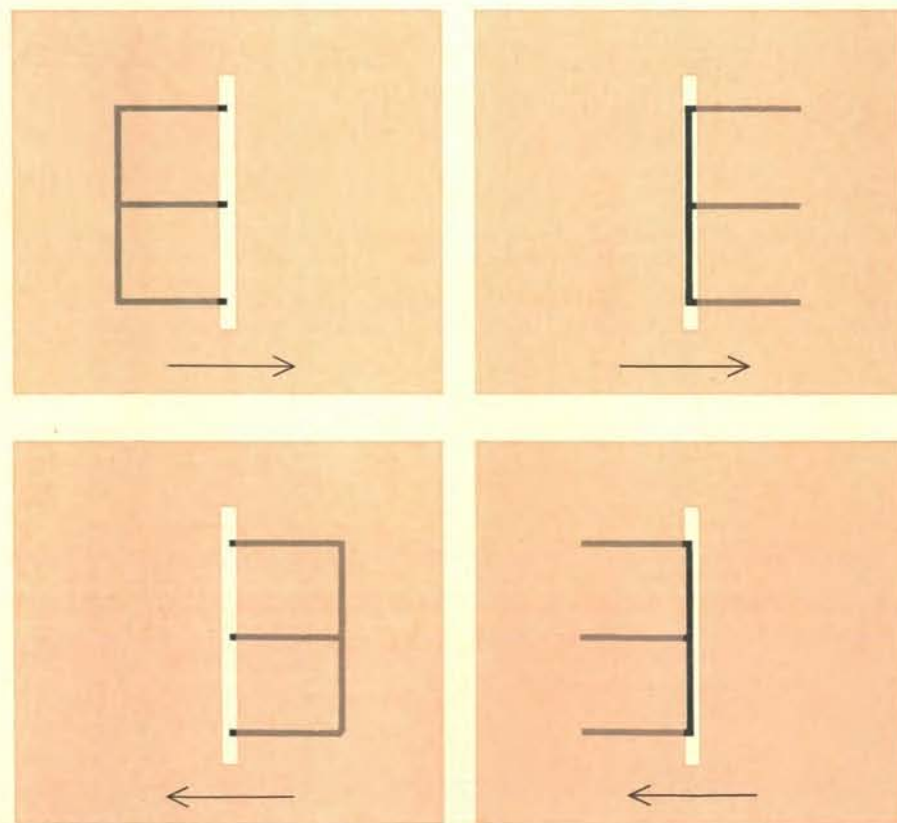
Abbiamo trovato un altro elemento importante. Se l'osservatore sa che ciò che gli viene mostrato può rappresentare una figura dietro una fenditura, è più probabile che la percepisca. Generalmente il semplice essere a conoscenza di qualcosa (per esempio sapere che un'illusione è un'illusione) non modifica ciò che si percepisce, ma in questo caso sì. Tuttavia anche con una conoscenza precedente, le condizioni devono essere favorevoli perché sia percepita una forma. Tale conoscenza non ha effetto se l'inclinazione del segmento non cambia mai, o se la figura è luminosa e vista al buio, in modo che la fenditura non sia visibile.

La nostra ricerca sulle caratteristiche di presentazione che regolano il risultato percettivo conduce al problema se l'effetto anortoscopico sia genuinamente percettivo, quando si è esclusa la pittura retinica. Molti ricercatori hanno sostenuto che, se in ciascun momento è visibile solo una parte della figura, l'impressione che si ha della figura completa dipende dal sapere che cosa è presente, piuttosto che essere il risultato di una vera percezione della forma. Soltanto se si vede la figura completa in un solo colpo, dicono, l'esperienza può essere descritta come percezione di forma.

Secondo me qui vi è una confusione che può essere ricondotta a un pregiudizio teorico. Se si crede che la percezione della forma presupponga un'immagine retinica estesa le cui parti siano tutte presenti simultaneamente, allora l'effetto anortoscopico non può essere di natura percettiva, se si esclude la pittura retinica. Tuttavia non c'è ragione di ricondurre la percezione della forma alla simultaneità. Sentire una melodia o una frase è certamente un fatto percettivo, anche se ciò che è percepito si estende nel tempo. Soltanto se la figura o la fenditura fossero mosse molto lentamente, sarebbe corretto dire che non si riesce a vedere la figura anche se si sa che c'è. Analogamente, si potrebbe impedire la percezione di una melodia separando i suoni con intervalli molto lunghi.

Il fatto è che quanti hanno sperimentato l'effetto anortoscopico in laboratorio sono sicuri che si tratti di un fenomeno percettivo. Una ragione di certezza sta sicuramente nelle chiare differenze osservate nelle diverse condizioni sperimentali. A questo punto gli insuccessi divengono molto importanti. Degli osservatori possono sapere che è stata presentata una figura, ma non riescono mai a percepirla a meno che le condizioni non lo permettano. In altre parole, ci sono dei casi in cui è appropriato dire che gli osservatori sanno che viene presentata una figura, ma non la percepiscono; questi casi, per contro, chiariscono come gli altri siano di natura percettiva. Tuttavia ci sono molte situazioni in cui la percezione slitta da una possibilità all'altra durante la presentazione; questa reversibilità in condizioni ambigue è una delle caratteristiche distintive del processo percettivo.

C'è una particolarità della percezione anortoscopica che non ho ancora citato. Quando la fenditura è immobile e la figura si muove, generalmente la figura percepita è distorta; tipicamente viene compressa lungo l'asse del suo movimento. Così un cerchio può apparire come un'ellisse, con l'asse maggiore verticale. Da cosa dipende questo effetto? I sostenitori dell'ipotesi della pittura retinica hanno suggerito che la distorsione sia conseguenza dell'incapacità dell'osservatore di muovere gli occhi in perfetta sincronia con la figura. Se il movimento dell'occhio non tiene testa alla velocità o all'ampiezza del movimento della figura, l'immagine «dipinta» sulla retina sarebbe compressa, e ciò renderebbe conto con facilità della distorsione.

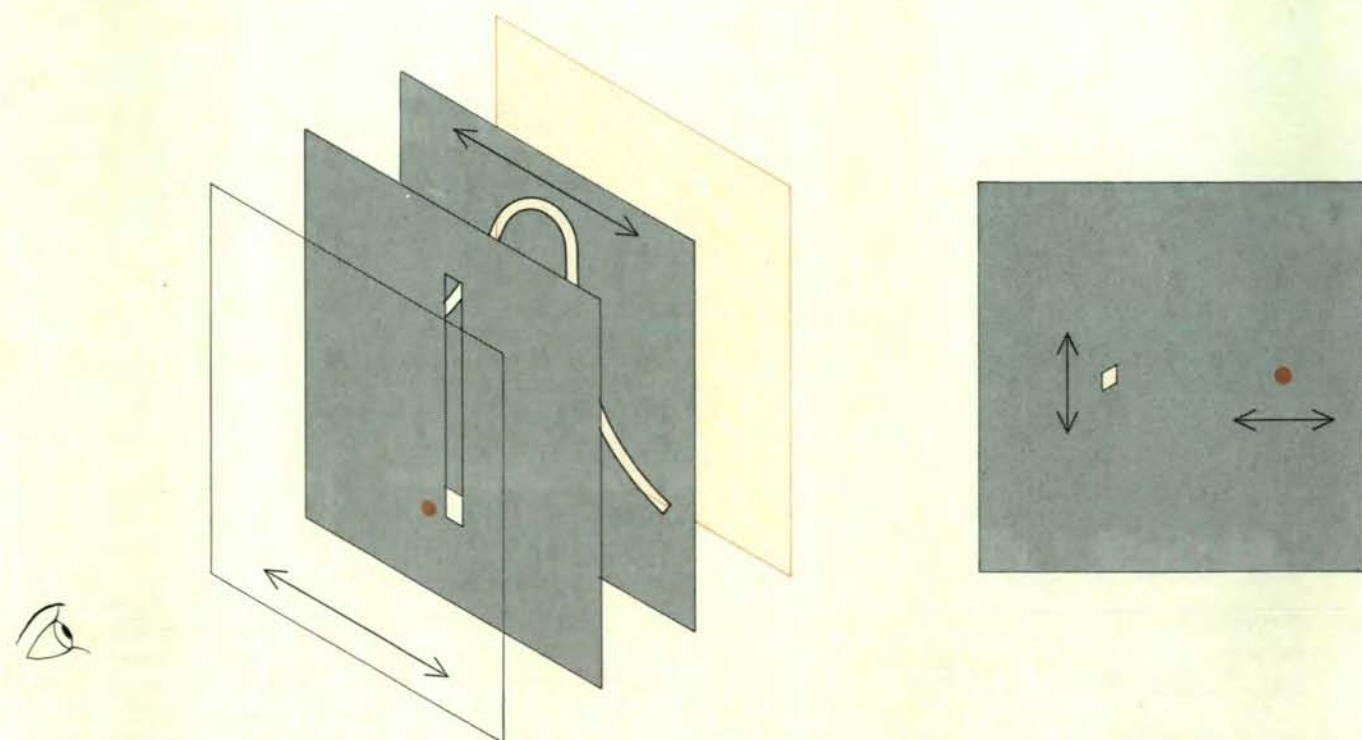
Un esperimento eseguito nel 1967 da Stuart M. Anstis e Janette Atkinson dell'Università di Bristol tende a confermare questa ipotesi. Essi avevano introdotto un bersaglio mobile (un punto luminoso) che l'osservatore doveva seguire avanti e indietro. Variando la velocità del bersaglio in relazione alla velocità della figura, Anstis e Atkinson erano in grado di «dipingere» un'immagine retinica della figura che variava nella forma. Per esempio, se la figura in movimento era un cerchio e il bersaglio da seguire si muoveva a velocità dimezzata, l'immagine diffusa sulla retina era un'ellisse, il cui asse verticale era lungo il doppio di quello orizzontale.

Se un osservatore giudica erroneamente la direzione del movimento di una figura dietro una fenditura immobile, la figura risulta ribaltata. Nella situazione illustrata appare la lettera *E* se la direzione del movimento è percepita correttamente, cioè da sinistra verso destra *(sopra)*; se invece è interpretata erroneamente da destra verso sinistra *(sotto)*, si vede la figura invertita Ǝ.

Se la figura si muoveva verso destra mentre il bersaglio da inseguire andava a sinistra, l'immagine prodotta era ribaltata rispetto a quella prodotta dalla normale percezione. Gli osservatori riferivano di vedere delle figure la cui forma corrispondeva esattamente all'immagine retinica stabilita.

Questa scoperta ha imbarazzato i miei colleghi e me per un po' di tempo, poiché avevamo buone ragioni per dubitare che la pittura retinica rendesse conto dell'effetto anortoscopico o di queste distorsioni percettive. Alla fine abbiamo prospettato una spiegazione alternativa che ci sembra si accordi meglio con i fatti. Per comprendere questa spiegazione bisogna in primo luogo considerare più attentamente la percezione della figura quando la fenditura è immobile e la figura in movimento. In accordo con l'ipotesi della pittura retinica, ciò che sarebbe visto è la figura completa all'interno di una larga apertura rettangolare: il rettangolo rappresenta l'immagine della fenditura, che sarebbe anch'essa raffigurata sulla retina. Questo non è quello che si vede. Ciò che si vede è una figura che si muove dietro una stretta fenditura.

Il problema è: che cosa determina la velocità percepita della figura? Data la strettezza della fenditura, l'informazione circa la velocità della figura difficilmente può essere convogliata con accuratezza per mezzo del breve passaggio dei componenti distinguibili del contorno attraverso la fenditura. D'altronde, con le figure da noi impiegate non ci sono componenti distinguibili se non alla fine. Perciò la velocità della figura è rappresentata perlomeno in maniera ambigua. La lunghezza percepita della figura dipende interamente dalla sua velocità percepita per lo meno in accordo con l'interpretazione della percezione anortoscopica, in termini di soluzione di problemi. Dal momento che il risultato più comune quando non c'è un bersaglio da inseguire è una compressione percepita, concludiamo che la velocità è sottostimata. Noi non sappiamo perché le cose stiano così, ma è importante tener presente che non c'è neanche ragione di aspettarsi un'accurata percezione della velocità.

Quando viene introdotto un bersaglio da inseguire, il sistema percettivo sembra assumere che la figura si sta muovendo alla velocità del bersaglio. Si può formulare un'ipotesi generale in base alla quale il movimento dell'occhio è un indizio sul movimento della figura nella situazione anortoscopica. Avuto questo indizio, la velocità percepita della figura è raddoppiata quando il bersaglio si muove a velocità doppia rispetto alla figura ed è dimezzata quando la figura si muove con velocità doppia rispetto al bersaglio. Poiché la lunghezza apparente della figura dipende dalla distanza a cui sembra muoversi dietro la fenditura durante l'intervallo fra la comparsa del primo estremo e la comparsa del secondo, la distorsione trovata da Anstis e Atkinson risulta spiegata. Essa deriva da una costruzione mentale della lunghezza, derivata dalla velocità apparente della figura e non direttamente da una immagine retinica distorta.

Per procurare un sostegno alla nostra interpretazione Di Vita, Halper, Deborah Wheeler e io abbiamo condotto un esperimento in cui gli osservatori vedevano una linea curva muoversi dietro una fenditura a una certa velocità e nel medesimo tempo inseguivano un punto bersaglio. In effetti, abbiamo ripetuto l'esperimento di Anstis e Atkinson, ma con il nostro tipo di figura. Inoltre vi era un'importante aggiunta nella procedura. L'osservatore non soltanto indicava la lunghezza percepita della figura (regolando un dispositivo che proiettava un'ombra che variava la lunghezza di una replica della figura mentre teneva costante la sua altezza), ma anche ci riferiva se la velocità della figura sembrava essere uguale, inferiore o maggiore di quella del bersaglio mobile. Sono state fatte osservazioni a differenti velocità del

L'autore e i suoi colleghi in uno dei loro esperimenti di controllo sulla percezione anortoscopica hanno impiegato un insieme di superfici multiple. La prima superficie davanti all'osservatore è costituita da un foglio trasparente di plastica su cui è montato un bersaglio luminoso da inseguire *(punto in colore)*; questo foglio si muove avanti e indietro. La superficie successiva è un cartoncino opaco immobile con una fenditura verticale. La terza superficie è un cartoncino opaco su cui è ritagliata la figura test; anche questa superficie si muove avanti e indietro. La superficie posteriore è un foglio di plastica traslucida che diffonde la luce proveniente dalla batteria di lampade posta sul retro. Il disegno sulla destra mostra ciò che vede l'osservatore quando l'esperimento si svolge in una camera oscura: un bersaglio luminoso, da inseguire, che si muove orizzontalmente e un segmento di linea che sembra muoversi verticalmente (o quasi verticalmente). In queste condizioni sperimentali non viene vista la figura in movimento dal momento che non viene percepita la fenditura. Quando, invece, si impiega la stessa presentazione, ma la figura sulla terza superficie è disegnata e la stanza è illuminata, l'osservatore vede in ogni caso una figura in movimento.

bersaglio: uguale alla velocità della figura, dimezzata, doppia rispetto a quella della figura, uguale a quella della figura, ma in direzione opposta. Avevamo anche una condizione di «visione libera», senza la presenza del bersaglio in movimento.

Quando seguivano il bersaglio, tutti gli osservatori percepivano una figura mentre in «visione libera» nessuno la percepiva. Per questo esperimento avevamo deliberatamente usato una fenditura molto stretta (0,15 centimetri), condizione che sapevamo avrebbe eliminato l'effetto anortoscopico nella più ordinaria situazione di «non inseguimento».

La prima cosa da rilevare nei risultati è che l'inseguimento è un fattore importante della percezione di una figura anortoscopica. Riteniamo che la ragione più importante di questo dato sia dovuta al fatto che il movimento degli occhi procura un indizio efficace circa il movimento avanti e indietro della figura dietro la fenditura. Inoltre, questo indizio specifica la velocità della figura, che altrimenti è indeterminata. Sarebbe difficile arrivare alla percezione di una figura di lunghezza definita se la velocità fosse indeterminata. Praticamente tutti gli osservatori hanno riferito che la velocità della figura era la stessa o per lo più la stessa del bersaglio senza riferimento alla loro velocità assoluta. Così per la lunghezza percepita della figura, se la lunghezza reale è posta uguale a 1, la lunghezza percepita valeva 0,77 quando le due velocità erano le stesse, 1,65 quando la velocità del bersaglio era doppia e 0,50 quando era la metà di quella della figura.

Quando la figura si muoveva in una direzione e il bersaglio in quella opposta, ma alla stessa velocità, tutti gli osservatori percepivano la figura come un'immagine speculare della sua vera forma. Per questa fase sperimentale abbiamo impiegato una figura asimmetrica e richiedevamo agli osservatori un rapido schizzo di quello che percepivano. Il risultato era quello che si poteva prevedere sulla base dell'immagine rovesciata raffigurata sulla retina, ma anche quello che si poteva prevedere sulla base della nostra ipotesi relativa agli indizi dei movimenti oculari. Non soltanto era ambigua la velocità della figura dietro la fenditura, lo era anche la sua direzione. Per esempio, se la figura era la lettera E e si muoveva verso destra, gli estremi dei suoi tre denti orizzontali si sarebbero mostrati per primi. Se la figura sembrava muoversi verso sinistra, ma i punti terminali dei denti apparivano ancora per primi ed erano seguiti dal contorno verticale della *E*, la figura avrebbe dovuto essere una figura ribaltata da destra a sinistra: Ǝ. Che la direzione del movimento fosse percepita erroneamente in questo modo è testimoniato dalle risposte errate di tutti quegli osservatori che percepivano la figura ribaltata quando si chiedeva loro quale fosse la direzione del suo movimento.

In questo ultimo esperimento l'inseguimento di un bersaglio che si muove ha l'effetto di diffondere l'immagine della figura sulla retina. Anche se abbiamo rifiutato l'ipotesi che l'immagine così «dipinta» sia la causa diretta della percezione anortoscopica della forma e delle distorsioni osservate, un aspetto dei risultati suggerisce un importante ruolo dell'immagine estesa. Il percetto della figura è più chiaro e più simile alla comune percezione della forma in tali condizioni che non quando gli occhi sono immobili mentre osservano una fenditura fissa. Abbiamo trovato un risultato simile quando, nella condizione di fenditura mobile, si introduce un punto fisso e l'osservatore tiene lo sguardo fermo su di esso. Anche qui l'immagine è propagata sulla retina, mentre, seguendo la fenditura che si muove, non abbiamo una tale immagine estesa. In entrambi i casi, quando è presente questa immagine, quasi sempre un osservatore riesce a percepire una figura, anche se le condizioni sono altrimenti insufficienti per l'effetto anortoscopico.

Chiamiamo effetto di facilitazione il miglioramento nella percezione anortoscopica che può essere attribuito a una immagine retinica estesa. Evidentemente il sistema percettivo trova più facile integrare le parti successive della figura in una forma completa quando le medesime vengono propagate attraverso la retina che non quando cadono tutte su una striscia verticale. Sebbene l'immagine estesa possa facilitare l'effetto anortoscopico, non lo provoca direttamente. I motivi di tale conclusione sono stati esposti in precedenza, e possono essere sostenuti da ulteriori esperimenti di controllo.

In uno di questi esperimenti la figura era una linea retta orizzontale e l'osservatore inseguiva un bersaglio che si muoveva alla stessa velocità della figura. Sebbene sulla retina venisse propagata un'immagine della figura, nessun osservatore ha percepito una figura. Al contrario, il segmento visibile della linea sembrava essere parte della fenditura e non si aveva affatto l'impressione di qualcosa che si muovesse (eccetto il punto bersaglio). Possiamo solamente concludere che senza alcuno spostamento verticale del contorno il sistema percettivo non ha motivo di dedurre che dietro la fenditura una figura estesa si sta muovendo; di conseguenza tale immagine estesa, sebbene presente, non è integrata in una forma costruita mentalmente.

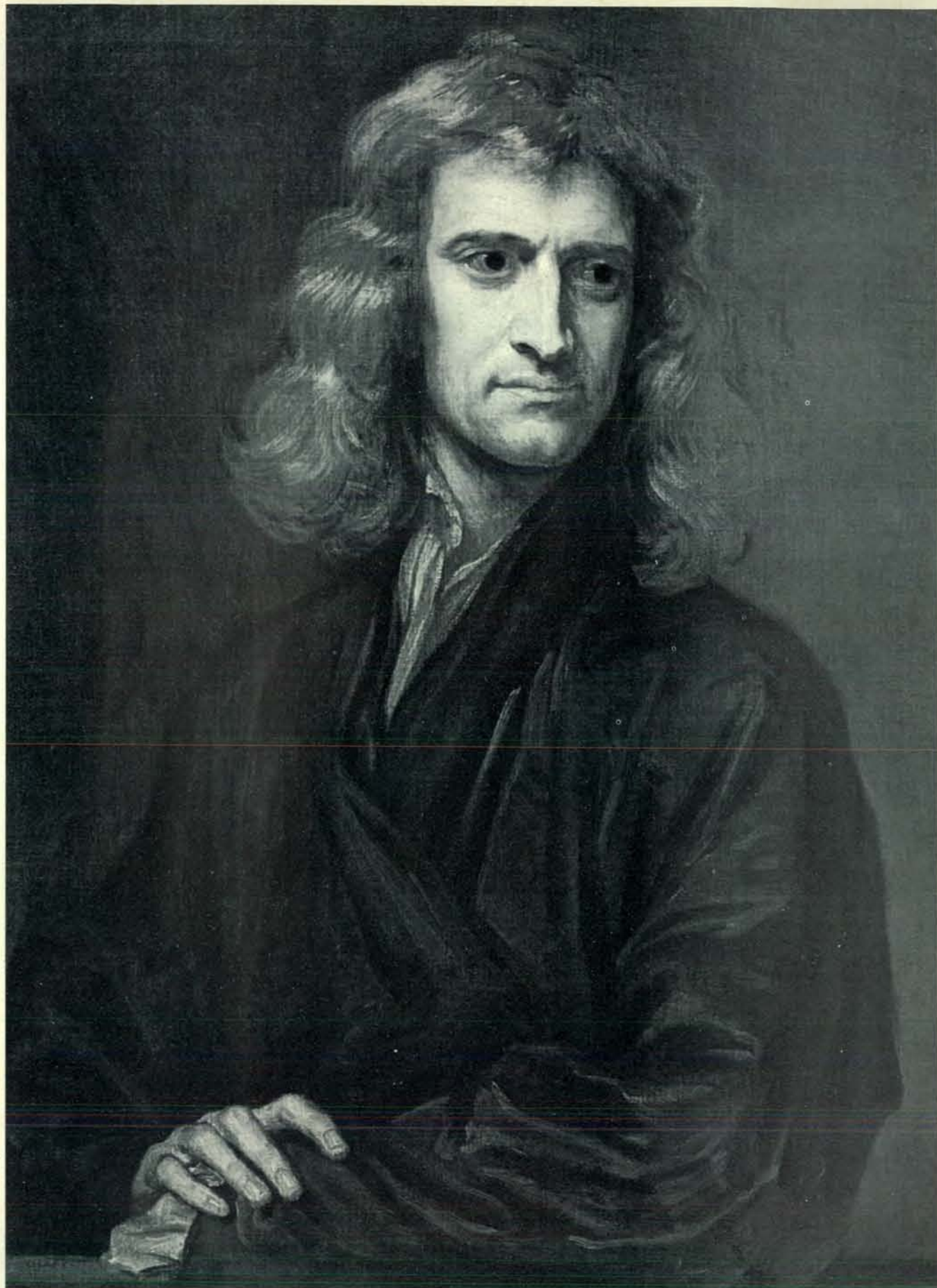
In un altro esperimento di controllo la figura consisteva in una riproduzione ritagliata di una figura curva, che era illuminata da dietro. C'era anche un punto luminoso, e la presentazione veniva osservata in una camera per il resto oscura. Quando l'osservatore seguiva il bersaglio mobile, l'immagine della linea luminosa veniva propagata sulla retina. Tuttavia non veniva percepita nessuna figura. Il segmento visibile della figura, che consisteva essenzialmente di un piccolo punto di luce, appariva piuttosto muoversi su e giù, sebbene la sua traiettoria sembrasse leggermente inclinata rispetto alla verticale.

Dal momento che l'inseguimento ha l'effetto di propagare l'immagine del frammento visibile sulla retina, il fatto che in questo caso non si riesca a ottenere un percetto della figura, dimostra decisamente come l'effetto anortoscopico dipenda dalla percezione della fenditura, in quanto apertura su una superficie opaca. Solamente allora esiste il sostegno necessario per la costruzione mentale di una figura estesa.

In ogni situazione sembrano presenti due fattori che contribuiscono alla buona riuscita della percezione anortoscopica della forma, quando viene inseguito un bersaglio mobile. Il movimento degli occhi serve come indizio che una figura si muove perpendicolarmente alla fenditura, e fornisce una velocità non ambigua alla figura stessa. La formazione di un'immagine diffusa sulla retina facilita l'integrazione della figura. Forse è proprio questa facilitazione che ha portato alcuni ricercatori a fare propria l'ipotesi della pittura retinica.

In conclusione ho sostenuto che la percezione della forma può essere compresa come un processo di integrazione di informazioni circa la posizione rispettiva delle parti di una figura (purché siano organizzate in una unità e siano interpretate come figura piuttosto che come sfondo). I contorni fisici o le loro rappresentazioni nell'immagine retinica non sono necessari, a patto che qualche tipo di informazione indichi dove sono i confini di una figura. Conseguentemente la presentazione anortoscopica di una figura, una parte alla volta, può sostenere la percezione della forma anche se una immagine estesa della figura completa non compare sulla retina in alcun istante.

Poiché la presentazione anortoscopica è ambigua e non rappresenta necessariamente una figura parzialmente coperta, ne consegue che la possibilità di una percezione di forma comporta un processo di soluzione di problemi. La presenza di una regione vista come una stretta apertura in una superficie opaca circostante è indispensabile, e lo stimolo deve possedere anche determinate altre proprietà, se la percezione di una figura deve essere la soluzione preferita. Nel caso in cui una figura si muova dietro una fenditura immobile il sistema percettivo deve anche inferire la velocità e la direzione della figura al fine di ricostruire la sua lunghezza e la sua forma.

Se questa interpretazione dei fatti che accompagnano la visione di una presentazione anortoscopica è corretta, la percezione della forma è un processo molto più vicino al livello cognitivo di quanto si sia pensato sino ad ora. Non può essere spiegata come il risultato diretto di una elaborazione fisiologica dei contorni che stimolano la retina.

Questo ritratto fu dipinto da Godfrey Kneller nel 1689, quando Newton aveva 46 anni. Quattro anni prima, Newton aveva sviluppato il concetto della gravitazione universale. L'opera principale di Newton, i *Philosophiae naturalis principia mathematica*, fu edita nel 1687.

Newton e la scoperta della gravità

In che modo Newton pervenne a sviluppare il concetto che segnò l'inizio della scienza moderna? In sostanza, attraverso una serie di confronti fra il mondo reale e una sua rappresentazione matematica semplificata

di I. Bernard Cohen

Il culmine della Rivoluzione scientifica si identifica con la scoperta della legge della gravitazione universale a opera di Newton: tutti gli oggetti si attraggono reciprocamente con una forza che è direttamente proporzionale al prodotto delle loro masse e inversamente proporzionale al quadrato della loro distanza. Riuscendo a trattare mediante una singola legge matematica i principali fenomeni fisici dell'universo osservabile, Newton dimostrò che fisica terrestre e fisica celeste sono una sola fisica. D'un sol colpo il concetto della gravitazione universale rivelò il significato fisico delle tre leggi dei moti planetari di Keplero, risolse lo spinoso problema dell'origine delle maree e rese ragione dell'osservazione curiosa e inspiegata di Galileo che la discesa di un corpo in caduta libera è indipendente dal suo peso. Newton aveva realizzato l'obiettivo di Keplero di sviluppare una fisica fondata su cause.

La fondamentale scoperta della gravitazione universale, che divenne il paradigma di una scienza trionfante, non fu il risultato di un lampo di genio isolato, ma il culmine di una serie di esercizi nella soluzione di problemi. Fu un prodotto non dell'induzione ma di deduzioni logiche e di trasformazioni di idee esistenti. La scoperta della gravitazione universale rivela quella che io credo sia una fra le caratteristiche principali di ogni progresso scientifico, dalle più semplici innovazioni alle rivoluzioni più drammatiche: la creazione di qualcosa di nuovo attraverso la trasformazione di nozioni esistenti.

Newton sviluppò il concetto di gravitazione universale nei primissimi mesi del 1685, a quarantadue anni. I fisici hanno dato di solito i loro massimi contributi in età molto più giovane, ma Newton si trovava ancora in quelli che definì «i primi anni della mia vita per l'invenzione». I documenti che mi hanno consentito di datare la scoperta rendono possibile anche una ricostruzione del processo che condusse a essa.

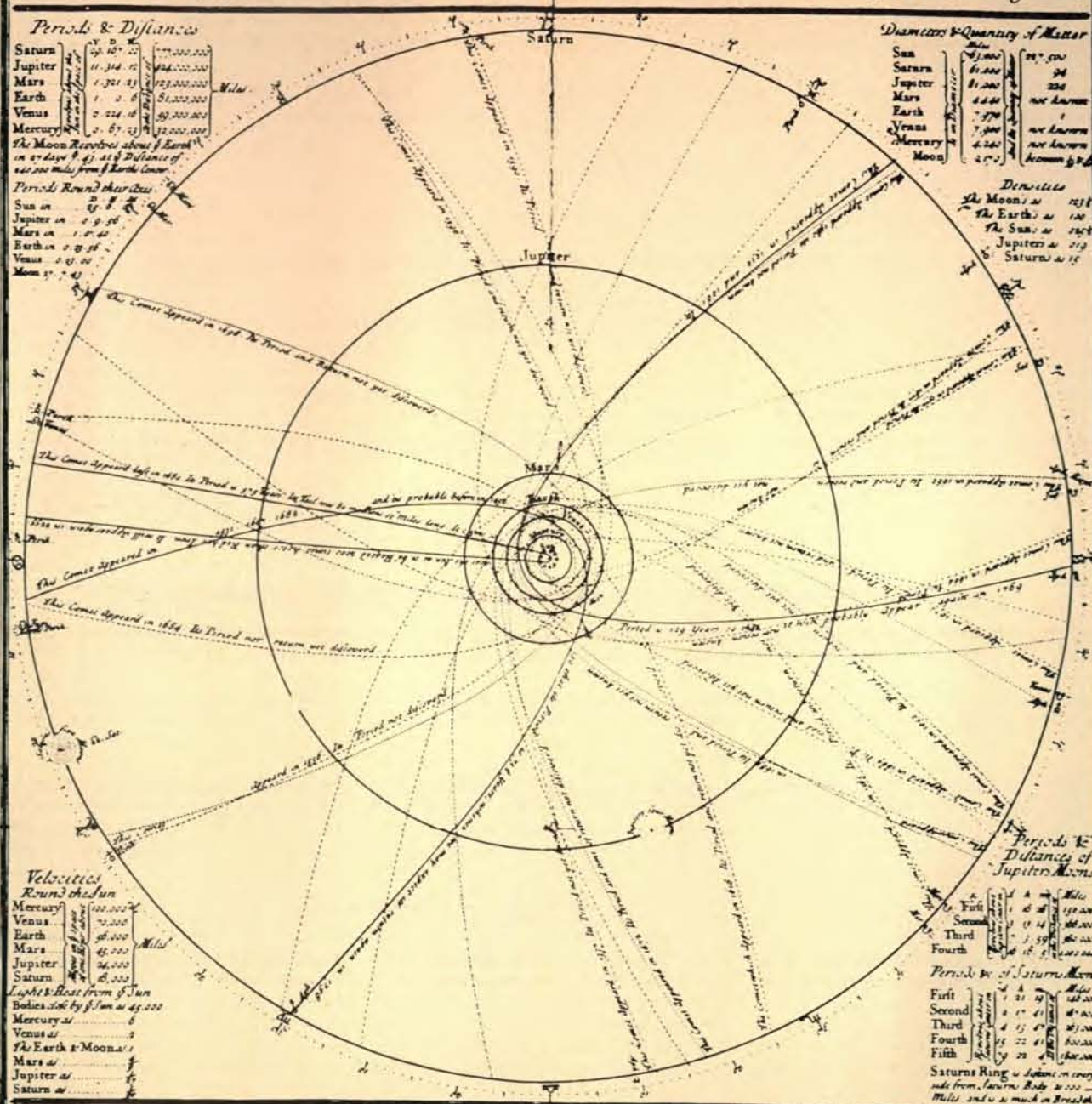
Un passo decisivo sulla via verso la gravitazione universale fu compiuto fra la fine del 1679 e l'inizio del 1680, quando Robert Hooke introdusse Newton a un modo nuovo di analizzare un moto lungo una traiettoria curva. Hooke si era reso conto che il moto di un corpo orbitante aveva due componenti, una componente inerziale e una componente centripeta, ossia diretta verso il centro. La componente inerziale tende a far muovere il corpo su una linea retta tangente all'orbita curva, mentre la componente centripeta lo attrae continuamente facendolo deviare dalla traiettoria rettilinea inerziale. In un'orbita stabile come quella della Luna le due componenti si fanno equilibrio, cosicché la Luna non si allontana lungo una traiettoria rettilinea né discende a spirale verso la Terra.

Il concetto di forza centripeta sostituì la nozione anteriore, e sviante, di forza centrifuga, in allontanamento dal centro. René Descartes e Christiaan Huygens avevano analizzato moti curvi in funzione di una tale forza centrifuga. Descartes, per esempio, aveva investigato il moto di una palla sulla superficie interna di un cilindro cavo e il movimento dell'acqua in un secchio mosso velocemente in cerchio. La palla e l'acqua sembravano fuggire dal centro del sistema, cosicché Descartes ne attribuì il moto all'influenza di una forza centrifuga. Ora era chiaro che una tale forza non esiste; una forza centrifuga non può essere ricondotta all'interazione fra oggetti fisici. L'illusione di una forza centrifuga si verifica quando un oggetto in movimento viene osservato da un sistema di riferimento in rotazione.

Col mutamento di prospettiva dalla forza centrifuga alla forza centripeta si pervenne a un apprezzamento della funzione fondamentale del corpo centrale. L'analisi centrifuga si era concentrata sull'oggetto in rotazione, il cui «sforzo per allontanarsi» dal centro sembra indipendente dalle proprietà del corpo centrale. Il concetto di una forza centripeta dipende invece fondamentalmente dal corpo centrale, verso cui l'oggetto in rivoluzione viene attratto.

L'analisi compiuta da Hooke del moto curvo potrebbe sembrare una conseguenza così ovvia e immediata del principio d'inerzia cartesiano da dar l'impressione che Newton non avesse certo bisogno che Hooke lo istruisse sull'argomento ancora nel 1679. Newton aveva infatti già più o meno accettato il principio inerziale una ventina di anni prima. Ma Newton, come Descartes e Huygens, era così impantanato nel concetto di forza centrifuga da avere tutt'altro che chiare le piene implicazioni della fisica inerziale.

Il 24 novembre 1679 Hooke scrisse a Newton proponendogli di avviare una corrispondenza «filosofica» privata su argomenti scientifici di reciproco interesse. Sei anni prima essi avevano avuto uno scontro pubblico sugli esperimenti e sulle teorie di Newton a proposito della dispersione della luce per mezzo di un prisma e della natura del colore. Hooke era solo uno fra i vari ricercatori che avevano rifiutato le teorie ottiche di Newton. Newton fu così irritato dal fatto di dover difendere la sua opera che giurò di abbandonare la «filosofia» (la scienza fisica), poiché essa era una «signora così litigiosa» che un uomo che avesse avuto qualcosa a che fare con lei avrebbe dovuto spendere il resto della sua vita a difendere le sue opinioni.

Hooke era poi diventato segretario della Royal Society di Londra. Nonostante la citata controversia sull'ottica, la sua lettera a Newton era amichevole e benevola. Invitava Newton a commentare qualsiasi delle sue ipotesi o opinioni,

THIS moſt Elegant Syſtem of the Planets and Comets could not be produced but by and under the Contrivance and Dominion of an Intelligent and Powerful Being. And if the Fixed Stars are the Centers of such other Syſtems, all theſe being Framed by the like Council will be Subject to the Dominion of *One*, eſpecially ſeeing the Light of the Fixed Stars is of the ſame Nature with that of the Sun, and the Light of all theſe Syſtems paſses mutually from one to another. He governs all things, not as ye Soul of the World, but as the Lord of the Univerſe, and becauſe of his Dominion he is wont to be called Lord God παντοκράτωρ (i.e. Univerſal Emperor) for God is a Relative word, and has Relation to Servants: And the Deity is the Empire of God, not over his own Body (as is the Opinion of thoſe, who make him the Soul of the World) but over his Servants. The *Supreme God* is a Being Eternal, Infinite Abſolutely Perfect; but a Being however Perfect without Dominion, is not *Lord God*: For we say, *my God, your God, the God of Iſrael*, but we do not ſay, *my Eternal, your Eternal, the Eternal of Iſrael*; we do not ſay *my Infinite, your Infinite, the Infinite of Iſrael*; we do not ſay *my Perfect, your Perfect, the Perfect of Iſrael*. Theſe Titles have no Relation to Servants. The word God frequently ſignifies Lord, but every Lord is not God. The Empire of a Spiritual being conſtitutes God, true Empire conſtitutes True God, Supreme the Supreme, Feigned the Feigned. And from his true Empire it follows that the true God is Living, Intelligent & Powerful, from his other Perfections that he is the Supreme or Supremly Perfect. He is *Eternal & Infinite, Omnipotent* and *Omnipreſent* that is, he endures from Eternity to Eternity, and he is preſent from Infinity to Infinity, he Governs all Things, and Knows all Things which are or which can be known. He is not Eternity or Infinity, but he is Eternal and Infinite, he is not Duration or Space, but he Endures and is Preſent. He endures always and is preſent everywhere and by exiſting always and everywhere, he Conſtitutes Duration and Space, Eternity and Infinity. Whereas every Particle of Space is *always*, and every Individual Moment of Duration is *every where*, certainly the Framer and Lord of the Univerſe ſhall not be (*nunquam nuſquam*) *never no where*. He is Omnipreſent not Virtually only, but alſo Subſtantially, for Power without Subſtance cannot Subſiſt. In him are contain'd and moved all things (ſo the Antients thought *) but without mutual Paſsion God ſuffers nothing from the Motions of Bodies. Nor do they ſuffer any Reſiſtance from the Omnipreſence of God. It is confeſs'd that the Supreme God exiſts Neceſſarily, and by the ſame Neceſsity he is *always* and *every where*. Whence alſo he is wholy Similar, all Eye all Ear, all Brain, all Arm, all the Power of Perceiving Underſtanding and Acting. But after a manner not at all Corporeal, after a manner not like that of Men, after a manner wholly to us unknown. As a Blind Man has no notion of Colours, so neither have we any notion of the manner how the moſt Wiſe God perceives and underſtands all things. He is wholly deſtitute of all Body and of all Bodily ſhape, and therefore cannot be ſeen, heard, nor touched, nor ought to be Worſhiped under the Repreſentation of any thing Corporeal. We have Ideas of his Attributes, but we know not at all what is the Subſtance of any thing whatever. We ſee only the Figures and Colours of Bodies, we hear only Sounds, we touch only the outward Surfaces, we ſmell only Odours and taſt Taſts, but we know not by any ſence or reflex Act the inward Subſtances, and much leſs have we any Notion of the Subſtance of God. We know him only by his Properties and Attributes and by the moſt Wiſe and Excellent Structure of things, and by Final Cauſes, but we Adore and Worſhip him on account of his Dominion. For God without Dominion, Providence & Final Cauſes is nothing elſe but Fate and Nature.

Engrav'd and Sold by John Senex at the Globe in Salisbury Court near Fleetstreet.

in particolare la nozione che i moti dei pianeti fossero composti da «un moto diretto per la tangente e un moto di attrazione verso il corpo centrale». Fu a quanto pare questa frase a richiamare l'attenzione di Newton sull'idea di decomporre un moto curvo in una componente inerziale e in una componente centripeta. Tuttavia non esiste alcuna prova del fatto che Newton avesse ancora raggiunto il livello di Hooke nella comprensione del moto circolare. In effetti Newton continuava spesso a parlare di moto orbitale nei termini di forza centrifuga.

Nella sua lettera Hooke azzardava il suggerimento che la forza centripeta che attrae un pianeta verso il Sole vari inversamente al quadrato della distanza. A questo punto Hooke era perplesso. Egli non riusciva a vedere le conseguenze dinamiche della sua profonda intuizione e perciò non era in grado di fare il salto dal livello dell'intuizione e della congettura a quello della scienza esatta. Non riusciva ad andare oltre perché gli mancavano sia la genialità matematica di Newton sia un apprezzamento della legge kepleriana delle aree, la quale figurò con grande rilievo nel successivo studio della dinamica celeste da parte di Newton. La legge delle aree dice che il raggio vettore che unisce il Sole a un pianeta descrive in tempi uguali aree uguali.

Il 28 novembre Newton rispose a Hooke dicendogli che prima di leggere la sua lettera del 24 non aveva mai «sentito parlare (a quanto posso ricordare) della sua ipotesi che i moti celesti dei pianeti siano composti da un moto diretto per la tangente alla curva» e da un moto «di attrazione» verso il Sole. Dopo avere ammesso che l'analisi di Hooke era nuova per lui, Newton cambiò immediatamente argomento passando a una sua fantasticheria: l'effetto della rotazione della Terra su un oggetto in caduta libera. Se un oggetto lasciato cadere potesse passare attraverso la Terra in rotazione, quale traiettoria descriverebbe? Newton concluse erroneamente che seguirebbe una traiettoria a spirale.

Nella successiva lettera di Hooke, datata 9 dicembre, questi colse l'errore di Newton e sottolineò che la traiettoria «assomiglierebbe a un'ellisse». Hooke era ansioso che Newton affrontasse il problema del moto dei pianeti, cosicché gli suggerì che la corretta descrizione di un oggetto in caduta attraverso la Terra e la sua propria analisi del moto planetario erano due casi di «moti circolari composti da un moto diretto e da un moto di attrazione verso un centro».

Il 13 dicembre 1679 Newton rispose cautamente alla correzione di Hooke, ma non fece commenti sull'analisi del moto circolare che gli veniva suggerita. Hooke non rinunciò. In una lettera scritta in data 6 gennaio 1680 tornò alla sua tesi sul moto curvo e ripeté la supposizione quantitativa che l'attrazione centripeta sia inversamente proporzionale al quadrato della distanza. Da questa supposizione Hooke concluse che la velocità del corpo in rivoluzione è inversamente proporzionale alla distanza dal centro. Egli sottolineò poi che la sua analisi «spiega in modo molto intelligibile e vero tutte le apparenze celesti». Newton non rispose a questa lettera.

Il 17 gennaio Hooke gli inviò una breve lettera supplementare in cui scrisse: «Ora rimangono da conoscere le proprietà di una linea curva (non circolare né concentrica) prodotta da una forza attrattiva centrale, la quale rende le velocità di discesa dalla linea tangente o da un moto rettilineo uniforme, a tutte le distanze, in proporzione duplicata reciprocamente presa». In terminologia moderna, il problema di Hooke può essere parafrasato nel modo seguente: se una forza di attrazione centrale costringe un oggetto a deviare dalla sua traiettoria inerziale e a muoversi in una curva, quale tipo di curva si ottiene se la forza di attrazione varia in proporzione inversa al quadrato della distanza? Hooke concluse: «Non ho dubbi sul fatto che, grazie al suo metodo eccellente, Lei troverà facilmente di quale curva debba trattarsi, e quali siano le sue proprietà, e suggerirà una ragione fisica di tale proporzione».

Newton, a quanto pare, pervenne quasi ad assolvere il compito assegnatogli da Hooke. Dimostrò che un'ellisse avrebbe soddisfatto le condizioni da lui poste. Nondimeno, non comunicò i risultati della sua dimostrazione né a Hooke né ad alcun altro sino all'agosto 1684, quando ricevette la visita dell'astronomo e matematico Edmund Halley. Halley si recò a far visita a Newton per chiedergli «quale pensava sarebbe stata la curva descritta dai pianeti nell'ipotesi che la forza di attrazione verso il Sole fosse reciproca al quadrato della loro distanza da esso». Il problema era stato lungamente discusso alla Royal Society. Halley e Christopher Wren furono incapaci di risolverlo, e Hooke non pubblicò mai una soluzione, pur sostenendo di essere riuscito a trovarla.

Udita la domanda, Newton rispose immediatamente: un'ellisse. Halley gli chiese come lo sapesse e Newton replicò: «L'ho calcolata». A quanto pare Newton non riuscì a trovare i calcoli, ma su richiesta di Halley li scrisse per la Royal Society nel piccolo trattato *De motu* (Sul moto). Nel *De motu* Newton descrisse le sue indagini sulla dinamica terrestre e celeste, ivi comprese le sue idee sul moto in uno spazio libero e in un mezzo resistente. Newton doveva aver già terminato il *De motu* il 10 dicembre 1684, poiché in quella data Halley disse alla Royal Society che Newton gli aveva mostrato recentemente quel curioso trattato.

L'esatta evoluzione delle idee di Newton nel periodo intercorso fra la sua corrispondenza con Hooke e il completamento della prima stesura del *De motu* non è documentata. Sono certo nondimeno che sia stato proprio il metodo di Hooke dell'analisi del moto curvo a mettere Newton sulla pista giusta. Anche se non tutti gli storici saranno d'accordo con me, io credo che il modo di affrontare la dinamica terrestre e celeste adottato da Newton nel *De motu*, e che egli sviluppò ulteriormente nella primavera seguente nel libro I dei *Philosophiae naturalis principia mathematica*, rappresenti il suo pensiero sulla dinamica planetaria quale gli era stato ispirato dalla corrispondenza epistolare con Hooke. In alcuni manoscritti autobiografici, Newton disse che la corrispondenza con Hooke fu anteriore o contemporanea alla sua dimostrazione (pubblicata per la prima volta nel *De motu* e poi nei *Principia*) che un oggetto che si muove di moto inerziale e che è simultaneamente soggetto a una forza centripeta la quale varia in ragione inversa al quadrato della distanza si muove lungo un'orbita ellittica.

Fu questa dimostrazione a rivelare il significato fisico della legge delle orbite ellittiche di Keplero (la legge dice che ciascun pianeta si muove su una traiettoria ellittica di cui il Sole occupa uno dei fuochi). Il lettore moderno potrebbe essere sorpreso nel sentire che non fu Keplero bensì Newton a rivelare la natura fondamentale delle leggi kepleriane dei moti planetari. Prima della pubblicazione dei *Principia* queste leggi (che erano chiamate addirittura ipotesi) non erano tenute nella considerazione di cui hanno invece goduto ininterrottamente dopo Newton.

La legge kepleriana delle aree, in particolare, non godette di molta considerazione nel Seicento. La maggior parte delle opere di astronomia non la citavano neppure. Per esempio, l'*Astronomia carolina* di Thomas Streete, da cui Newton copiò la terza legge di Keplero (secondo la quale il cubo della distanza media di un pianeta dal Sole è proporzionale al quadrato del suo periodo orbitale), non parla mai della legge delle aree e neppure accenna alla sua esistenza. La maggior parte degli astronomi del Seicento calcolavano le posizioni dei pianeti non con la legge delle aree, ma con una costruzione fondata su un vettore in rotazione uniforme emanante dal fuoco vuoto dell'orbita ellittica del pianeta *(si veda l'illustra-*

Questo diagramma del sistema newtoniano del mondo fu eseguito da William Whiston, che successe a Newton come professore lucasiano all'Università di Cambridge. Il diagramma fu stampato su un foglio a sé, edito nel 1724. Vi si osservano i pianeti e i satelliti di Giove e di Saturno, orbitanti attorno al Sole sotto l'azione della gravitazione universale. È degno di nota il fatto che Whiston incluse nel diagramma anche le orbite di numerose comete. Newton aveva dimostrato che le orbite delle comete sono ellissi o parabole nelle quali il vettore che unisce il Sole alla cometa descrive in tempi uguali aree uguali. Sotto il diagramma vi è la traduzione inglese di Shiston di una parte dello Scolio generale finale dei *Principia* (tratto dalla seconda edizione del 1713). Qui Newton scrisse che «questo elegantissimo sistema dei pianeti e delle comete non poteva essere prodotto se non per opera del disegno e sotto il dominio di un essere intelligente e potente».

zione in alto a pagina 132). Poiché gli astronomi usavano raramente la legge delle aree, Newton dovette avere una percezione straordinaria per scorgerne il significato. Egli fu l'unico a elevare la legge delle aree di Keplero alla considerazione di cui essa gode tuttora.

La primissima proposizione dei *Principia* (e la discussione contenuta all'inizio del *De motu*) sviluppano il significato dinamico della legge delle aree dimostrando che il moto curvo descritto dalla legge è una conseguenza della forza centripeta. La dimostrazione, che è suddivisa in tre parti, ci mostra quanto bene Newton avesse appreso la tecnica di Hooke della scomposizione del moto curvo in una componente inerziale e una componente centripeta.

Nella prima parte della dimostrazione, Newton considera un corpo in moto lungo una linea retta con velocità costante. La linea è divisa in intervalli uguali per indicare che il corpo percorre in tempi uguali distanze uguali. A una distanza *h* al di sopra della linea del moto è scelto un punto *P*. I triangoli formati unendo il punto *P* con gli estremi di ciascuno degli intervalli uguali hanno la stessa area, essendo tutti triangoli di ugual base e di uguale altezza *h*. Con questa semplice analisi Newton rivelò l'esistenza di una relazione inattesa fra il moto inerziale e la legge delle aree.

Nella seconda parte della dimostrazione il corpo si muove inizialmente come prima, ma alla fine del secondo intervallo riceve una forza impulsiva - una spinta - verso *P*. Perciò nel terzo intervallo il corpo non si muove più lungo la linea retta originaria, ma lungo un'altra linea retta più vicina a *P*. Newton dimostrò di nuovo geometricamente che il triangolo formato unendo *P* agli estremi del tratto percorso nel terzo intervallo ha la stessa area del triangolo formato collegando *P* agli estremi del tratto percorso nel secondo intervallo.

Nella terza parte il corpo riceve un impulso verso *P* alla fine di ciascun intervallo. Di conseguenza viene a muoversi su una traiettoria poligonale attorno a *P*. In ogni intervallo continua a valere la legge delle aree. Nel caso limite in cui l'intervallo fra due impulsi consecutivi si approssimi a zero, il corpo è soggetto a una forza continua diretta verso *P* e la traiettoria poligonale diventa una curva o un'orbita continua. In questo modo Newton dimostrò che una forza centripeta genera una curva che obbedisce alla legge delle aree. La seconda proposizione dei *Principia* dimostra l'inverso: il moto lungo una curva che obbedisce alla legge delle aree implica una forza centripeta. Con queste due proposizioni Newton dimostrò che la legge delle aree è una condizione necessaria e sufficiente per il moto inerziale in un campo di forza centrale.

Le due proposizioni fanno parte di una serie di dimostrazioni che comincia con la legge delle aree e si conclude con una dimostrazione del fatto che un'orbita ellittica richiede una forza centripeta che obbedisca alla legge della proporzionalità inversa alla distanza. Questa serie di dimostrazioni, presentata sia nei *Principia* sia nel *De motu,* segna una profonda discontinuità nella storia delle scienze esatte. Queste dimostrazioni introdussero infatti una dinamica celeste radicalmente nuova fondata su nuovi concetti di forza, di quantità di moto, di massa e di inerzia e su una misura quantitativa totalmente nuova della forza dinamica. Il sottotitolo dell'*Astronomia nova* di Keplero proponeva l'obiettivo di creare una «fisica celeste fondata su cause». Newton realizzò questo obiettivo, di cui Keplero aveva avuto solo un'intuizione. Né Galileo né Descartes avevano concepito una tale dinamica celeste. E la formulazione newtoniana lasciò molto indietro anche il grande fisico Huygens.

Questa lettera, scritta da Robert Hooke a Newton, contiene l'esposizione delle concezioni di Hooke sull'analisi del moto lungo una traiettoria curva. (La lettera è datata 6 gennaio 1679 secondo una versione del calendario giuliano in cui l'anno cominciava in marzo; secondo il calendario moderno era effettivamente il 6 gennaio 1680.) Nella seconda frase Hooke suggerisce che «l'attrazione è sempre in una proporzione duplicata al reciproco della distanza dal centro» (ossia, l'attrazione è inversamente proporzionale al quadrato della distanza). Di conseguenza «la velocità sarà in una proporzione subduplicata all'attrazione e conseguentemente, come suppone Keplero, reciproca alla distanza». Hooke afferma che quest'analisi spiega «tutte le apparenze del cielo». Egli sottolinea l'importanza di «trovare le proprietà» delle curve perché dal moto curvo della Luna potrebbero venire derivate le longitudini, che sono «di grande importanza per l'umanità».

Dalla prima stesura del *De motu,* scritta probabilmente nel novembre 1684, è chiaro che Newton non aveva ancora sviluppato il concetto di gravita-

La forza centrifuga è una forza fittizia. L'illusione di una tale forza può originarsi quando un oggetto in movimento viene osservato da un sistema di riferimento in rotazione *(a sinistra)*, come quando una sfera viene tenuta appesa all'estremità di una cordicella da un osservatore che ruoti con la stessa velocità angolare della sfera stessa. Sulla sfera agiscono in questo caso due forze note: la tensione della funicella e la forza di gravità. La sfera non viene accelerata nella direzione della verticale, cosicché tutte le forze verticali che agiscono su di essa devono essere in equilibrio; in particolare, la componente verticale della tensione annulla la forza di gravità. Poiché l'osservatore e la sfera partecipano dello stesso movimento di rotazione, la sfera appare in quiete, e pare che anche le forze orizzontali dovrebbero essere in equilibrio. Di conseguenza l'osservatore postula una forza centrifuga che annulli la componente orizzontale della tensione. Nessuna forza del genere può nondimeno essere ricondotta all'interazione di oggetti fisici. Si ha un'analisi diversa delle forze *(a destra)* quando la sfera viene fatta ruotare nello stesso modo, ma l'osservatore è in quiete. In questo sistema di riferimento stazionario l'osservatore vede agire sulla sfera le stesse forze verticali che vedeva nel sistema di riferimento in rotazione. Nella direzione orizzontale la sfera non è invece in quiete rispetto all'osservatore, ma si muove in cerchio. La sfera è dunque soggetta a una forza centripeta, che è la componente orizzontale della tensione della funicella.

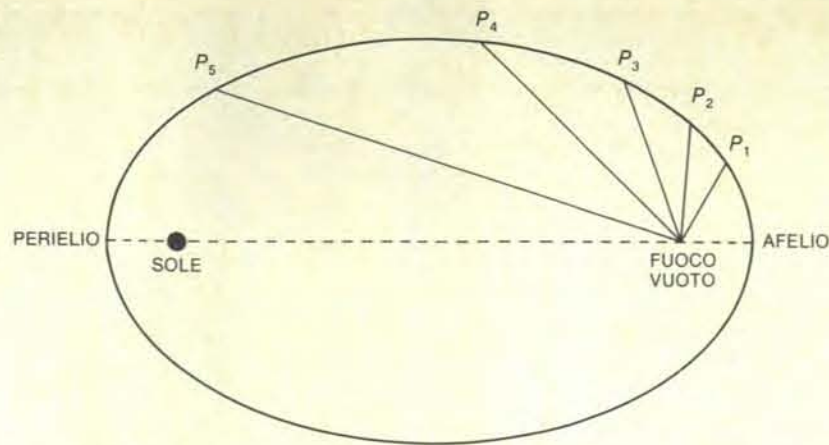
zione universale. Vi si discute infatti la forza centripeta diretta verso il fuoco di un'ellisse e vi si conclude con lo scolio seguente: «Perciò i pianeti maggiori orbitano in ellissi aventi un fuoco nel centro del Sole, e i raggi tirati [dai pianeti] al Sole descrivono aree proporzionali ai tempi, esattamente come suppose Keplero...».

Newton non dimostrò questo scolio, né continuò a credere in esso per molto tempo, e a rigore esso è falso. Com'egli si rese conto ben presto, i pianeti non si muovono secondo la legge delle aree in orbite ellittiche kepleriane semplici con il Sole in uno dei fuochi. Il fuoco si trova invece nel comune centro di massa, poiché non è soltanto il Sole ad attrarre ciascun pianeta, ma anche ciascun pianeta ad attrarre il Sole (e ad attrarsi reciprocamente). Se Newton a quest'epoca avesse già formulato il suo principio della gravitazione universale, non avrebbe proposto questo scolio erroneo.

Newton si rese conto rapidamente di non aver dimostrato che i pianeti si muovono precisamente secondo la legge delle orbite ellittiche e secondo la legge delle aree. Egli aveva trovato solo che le leggi valgono per un sistema a un corpo solo: una singola massa puntiforme in moto con una componente iniziale di moto inerziale in un campo di forza centrale. Riconobbe che il sistema a un corpo non corrisponde al mondo reale, ma a una situazione artificiale che è più facile da investigare matematicamente. Il sistema a un corpo riduce la Terra a una massa puntiforme e il Sole a un centro di forza immobile.

Ciò che consentì a Newton di andare oltre il sistema a un corpo fu la presa di coscienza delle conseguenze della sua terza legge del moto: la legge di azione e reazione. Questa legge è forse la più originale delle sue tre leggi del moto (le altre due sono la legge di inerzia e la legge della forza). Un indizio della sua novità può essere visto nel fatto che ancor oggi essa viene spesso usata in modo sbagliato da chi la riferisce non a una situazione di impatto o all'interazione fra corpi, ma a una supposta condizione di equilibrio.

Lo sviluppo del pensiero di Newton sull'azione e reazione dopo che egli ebbe completato la prima stesura del *De motu* è esposta nelle sezioni di apertura del libro I dei *Principia*. Nell'introduzione alla sezione XI Newton spiega che fino ad allora si era limitato a una situazione «pressoché inesistente nel mondo reale», ossia ai «moti di corpi attratti verso un centro immobile». Tale situazione era artificiale perché «le attrazioni sono dirette abitualmente verso corpi e - per la terza legge del moto - le azioni di corpi che attraggono e che sono attratti sono sempre reciproche e uguali». Perciò, «se ci sono due corpi, né il corpo che attrae né il corpo che è attratto possono essere in quiete». È vero piuttosto che «entrambi i corpi (per il quarto corollario delle leggi) orbitano attorno a un centro comune, come in virtù di una reciproca attrazione».

Newton si era reso conto che, se il Sole esercita un'attrazione sulla Terra, anche

Le posizioni dei pianeti venivano spesso determinate nel Seicento non per mezzo di un'applicazione della legge delle aree di Keplero, ma con una costruzione fondata su un raggio vettore in rotazione uniforme emanante dal fuoco vuoto dell'orbita ellittica di un pianeta. La posizione di un pianeta in momenti successivi (P_1, P_2, P_3, P_4, P_5) è l'intersezione dell'ellisse e del raggio vettore. La legge delle aree formulata da Keplero dice che il raggio vettore che unisce il Sole a un pianeta descrive in tempi uguali aree uguali. Di conseguenza il pianeta è più lento all'afelio che non al perielio. La costruzione qui rappresentata dà lo stesso risultato. Per rendere più esatta la concordanza della costruzione del grafico ai dati sono stati aggiunti alcuni fattori di correzione.

la Terra deve esercitare un'attrazione sul Sole con una forza di ugual grandezza. In questo sistema di due corpi, la Terra non si muove in un'orbita semplice attorno al Sole. Il Sole e la Terra si muovono invece attorno al loro comune centro di gravità. Un'altra conseguenza della terza legge del moto è che ciascun pianeta è, oltre che un corpo attratto, anche un centro di forza di attrazione; ne segue che un pianeta non solo attrae ed è attratto dal Sole, ma attrae anche ciascuno degli altri pianeti e ne è a sua volta attratto. Qui Newton ha compiuto un fondamentale passo avanti, passando da un sistema di due corpi interagenti a un sistema costituito da molti corpi interagenti.

Nel dicembre 1684 Newton completò una stesura riveduta del *De motu* nella quale i moti dei pianeti sono descritti nel contesto di un sistema di molti corpi interagenti. A differenza della stesura anteriore, quella riveduta conclude che «i pianeti non si muovono esattamente su ellissi né percorrono due volte la stessa orbita». Questa conclusione condusse Newton al risultato seguente: «Ci sono tante orbite di un pianeta quante sono le sue rivoluzioni, come nel caso della Luna, e l'orbita di ciascun pianeta dipende dal moto combinato di tutti i pianeti, per non menzionare le azioni che essi esercitano reciprocamente l'uno sull'altro». Newton aggiunse poi: «Una considerazione simultanea di tutte queste cause di moto e una definizione di questi moti per mezzo di leggi esatte che ammettano calcoli convenienti superano, se non sono in errore, la forza dell'intero intelletto umano».

Non c'è alcun documento che indichi come, nel mese circa intercorso fra la prima stesura del *De motu* e la sua revisione, Newton sia pervenuto a percepire che i pianeti agiscono gravitazionalmente l'uno sull'altro. Il passo citato sopra esprime nondimeno questa presa di coscienza in un linguaggio privo di ambiguità: «*eorum omnium actiones in se invicem*» (le azioni che essi esercitano reciprocamente l'uno sull'altro). Una conseguenza di quest'attrazione gravitazionale reciproca è che nessuna delle leggi di Keplero è valida in modo rigoroso nel mondo della fisica ma tutt'e tre sono valide solo per una costruzione matematica in cui masse puntiformi che non interagiscono fra loro orbitino o attorno a un centro di forza matematico o attorno a un corpo immobile che eserciti una forza di attrazione. La distinzione stabilita da Newton fra il regno della matematica, in cui le leggi di Keplero sono veramente leggi, e il regno della fisica, in cui esse sono solo «ipotesi», o approssimazioni, è uno dei caratteri rivoluzionari della dinamica celeste newtoniana.

Ho supposto che la terza legge del moto sia stata il fattore chiave nel ragionamento che condusse Newton a suggerire l'esistenza di perturbazioni gravitazionali reciproche delle orbite planetarie. A favore del mio assunto non c'è alcuna prova diretta, in quanto non esiste alcun documento che ci dia una versione anteriore dell'espressione «le azioni che essi esercitano reciprocamente l'uno sull'altro». Esiste nondimeno un probante indizio indiretto. Nella primavera del 1685, alcuni mesi dopo la revisione del *De motu*, Newton era ben avviato verso la conclusione della prima stesura dei *Principia*. Nella versione iniziale di quello che sarebbe diventato il libro II, *Sistema del mondo*, specificò i vari stadi che lo avevano condotto al concetto di interazioni gravitazionali planetarie. In questi stadi la terza legge del moto ha il ruolo principale e io non vedo alcuna ragione per credere che essi non siano gli stessi stadi che lo avevano guidato alcuni mesi prima, quando aveva rivisto il *De motu*.

Ecco due passi della prima stesura del *Sistema del mondo* che rivelano la funzione cruciale della terza legge del moto (trad. di Marcella Renzoni, Boringhieri, Torino, 1959, pp. 46-48):

«20. *L'accordo fra le analogie*.

«Poiché inoltre l'azione della forza centripeta sul corpo attratto è proporzionale, a pari distanza, alla materia contenuta in questo corpo, è ragionevole che essa sia proporzionale anche alla materia contenuta nel corpo che attrae. L'azione è perciò reciproca, e fa sì che i due corpi con un reciproco sforzo (per la terza legge del moto) si avvicinino vicendevolmente; tale azione deve perciò essere simile a se stessa in entrambi. Un corpo può essere considerato come quello che attrae, mentre l'altro è l'attratto: ma si tratta di una distinzione più matematica che naturale. L'attrazione in verità è esercitata da ognuno dei due corpi sull'altro, e pertanto è della medesima specie in entrambi.

«21. *E loro coincidenza*.

«Si può dunque affermare che la forza attrattiva si trova in entrambi. Il Sole attrae Giove e gli altri pianeti, Giove attrae i suoi satelliti, e per una eguale ragione questi agiscono l'uno sull'altro e su Giove, e tutti i pianeti agiscono reciprocamente gli uni sugli altri. E benché le azioni di due pianeti possano venire distinte l'una dall'altra, ed essere considerate come due azioni, mediante le quali ciascuno dei due attrae l'altro, tuttavia, in quanto sono intermedie [in quanto si esercitano fra i due corpi], non sono due, bensì costituiscono un'unica operazione fra due termini. Mediante la contrazione di una funicella che li unisca, due corpi possono attrarsi vicendevolmente. La causa dell'azione è duplice, cioè la disposizione dell'uno e dell'altro corpo; anche l'azione è duplice, in quanto si verifica in due corpi, ma, in quanto si verifica fra due corpi, è semplice ed unica. Per esempio, non abbiamo un'operazione con la quale il Sole attrae Giove, e un'altra operazione con la quale Giove attrae il Sole: si tratta in realtà di una sola operazione, con la quale il Sole e Giove cercano di avvicinarsi a vicenda. Mediante l'azione con la quale il Sole attrae Giove, Giove e il Sole (per la terza legge del movimento) si sforzano di avvicinarsi reciprocamente; e del pari, attraverso l'azione con la quale Giove attrae il Sole, Giove e il Sole tentano ugualmente di avvicinarsi. Ma il Sole non è attratto verso Giove mediante una duplice azione, e neppure Giove è attratto verso il Sole mediante una duplice azione, perché unica è l'azione intercorrente, con la quale ambedue si avvicinano.»

Newton concluse poi che «tutti i corpi si debbono attrarre reciprocamente secondo questa legge». Egli presentò orgogliosamente questa conclusione e spiegò perché l'entità della forza d'at-

La forza centripeta genera una traiettoria curva in accordo con la legge delle aree. Questa proprietà di una forza centripeta fu dimostrata da Newton nella prima proposizione dei *Principia* e nella discussione all'inizio del trattatello *De motu*. Newton considerò inizialmente *(a sinistra)* un corpo in moto lungo una linea retta con velocità costante. Il corpo comincia a muoversi in A_0 e dopo successivi intervalli uguali raggiunge dapprima A_1, poi A_2 e così via. Al di sopra della linea del moto viene scelto un punto *P*. I triangoli $A_0 PA_1$, $A_1 PA_2$, $A_2 PA_3$ e così via hanno tutti la stessa area, avendo basi uguali e la stessa altezza. Nella seconda fase dell'analisi *(al centro)*, il corpo comincia a muoversi come in precedenza, ma in A_2 riceve un impulso in direzione di *P*. Ora il corpo si muove lungo una linea retta diretta non più verso A_3 ma verso B_3. Newton dimostrò geometricamente che i triangoli $A_1 PA_2$ e $A_2 PB_3$ hanno la stessa area. Se alla fine di ciascun intervallo il corpo riceve un impulso verso *P (a destra)*, si muove in una traiettoria poligonale attorno a *P*. Anche in questo caso si formano triangoli aventi la stessa area. Nel caso limite in cui il tempo fra impulsi successivi si approssima a zero, il corpo è soggetto a una forza centripeta continua diretta verso *P* e la traiettoria poligonale diventa una curva continua. Anche in questo caso vale l'uguaglianza delle aree. Questa dimostrazione rivelò il significato dinamico della legge delle aree.

Pagina autografa del *De motu*, piccolo trattato scritto da Newton probabilmente nel novembre 1684. Nel *De motu* Newton si occupò di dinamica terrestre e dinamica celeste, compresa l'idea di una forza centripeta diretta verso il fuoco di un'ellisse. La pagina si conclude con lo scolio seguente: «Perciò i pianeti maggiori orbitano in ellissi aventi un fuoco *(umbilicum)* nel centro del Sole, e i raggi tirati (dai pianeti) al Sole descrivono aree proporzionali ai tempi, esattamente come suppose Keplero...». Lo scolio è sbagliato, e la natura dell'errore indica che Newton non aveva ancora sviluppato il concetto di gravitazione universale. Come Newton avrebbe poi compreso, il fuoco delle orbite planetarie non è il Sole, ma il centro di massa comune ai pianeti e al Sole. Non vi è solo l'attrazione esercitata dal Sole sui pianeti, ma anche quella esercitata dai pianeti sul Sole.

trazione sia così piccola da risultare inosservabile. «Si possono avvertire queste forze - scrisse - solo nei grandi corpi dei pianeti» (trad. it. cit., p. 49).

Nel libro III dei *Principia*, che si occupa anch'esso del sistema del mondo, ma che è un po' più matematico, Newton tratta l'argomento della gravitazione sostanzialmente nello stesso modo. Innanzitutto, nel caso della Luna, estende la forza peso, o gravità terrestre, sino alla Luna e dimostra che la forza varia in proporzione inversa al quadrato della distanza. Poi identifica la stessa forza terrestre con la forza esercitata dal Sole sui pianeti e con la forza esercitata da un pianeta sui suoi satelliti. Newton chiama ora tutte queste forze «gravità». Con l'aiuto della terza legge del moto trasforma il concetto di una forza esercitata dal Sole sui pianeti nel concetto di una forza reciproca che si esercita fra il Sole e i pianeti. Similmente, trasforma il concetto di una forza planetaria sui satelliti nel concetto di una forza reciproca fra pianeti e loro satelliti e fra satelliti.

La trasformazione finale è la nozione che tutti i corpi interagiscono gravitazionalmente fra loro.

La mia analisi dei vari stadi del pensiero di Newton non dovrebbe essere intesa come un tentativo di sminuire la forza straordinaria del suo genio creativo, ma piuttosto come un tentativo di renderne il genio plausibile. L'analisi illustra il fecondo modo di pensare la fisica proprio di Newton, nel quale la matematica viene applicata al mondo esterno quale è rivelato dall'esperimento e dall'osservazione critica. Questo modo di pensare, che sono solito chiamare stile newtoniano, è chiaramente espresso nel titolo del capolavoro di Newton: *Principi matematici di filosofia naturale*.

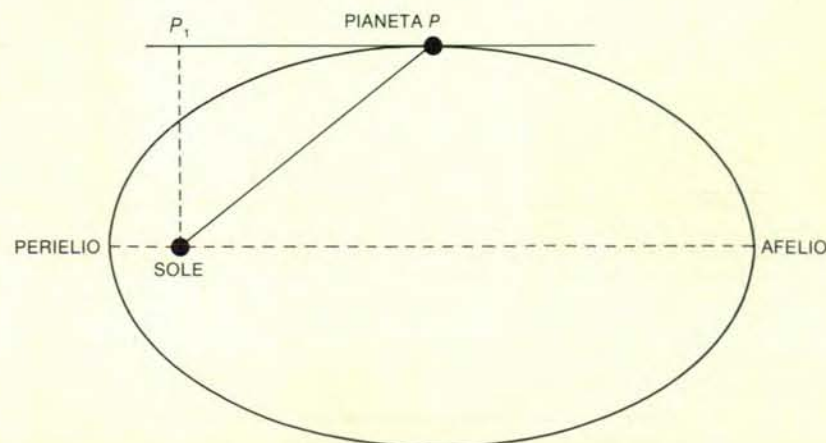
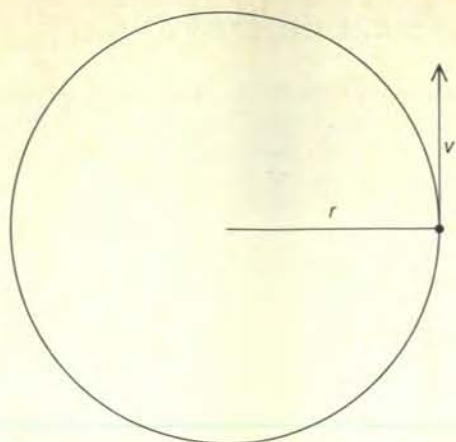
Lo stile newtoniano consiste in uno scambio ripetuto fra un costrutto matematico e la realtà fisica. Nello sviluppo delle idee di Newton sulla gravità e nella sua presentazione di tali idee nei *Principia*, egli prese l'avvio da un costrutto matematico che rappresenta una semplificazione della natura: una massa puntiforme in movimento attorno a un centro di forza. Non supponendo che il costrutto fosse una rappresentazione esatta dell'universo fisico, egli fu libero di esplorare le proprietà e gli effetti di una forza d'attrazione matematica, anche se il concetto di una forza di attrazione «agente a distanza» era ripugnante e inammissibile nell'ambito di una buona fisica. Poi confrontò le conseguenze del suo costrutto matematico con i principi e le leggi del mondo esterno desunti dall'osservazione, come le leggi kepleriane delle aree e delle orbite ellittiche. Quando il costrutto matematico si rivelava insufficiente, Newton lo modificava. Identificò il centro di forza non con un'entità matematica, ma con una massa puntiforme. Dico una massa puntiforme anziché un corpo fisico perché egli non aveva ancora considerato proprietà fisiche come volume, figura e massa.

Dal costrutto matematico modificato Newton concluse che un insieme di masse puntiformi orbitanti attorno alla massa puntiforme centrale si attraggono fra loro e perturbano a vicenda le loro orbite. Poi confrontò di nuovo il costrutto col mondo fisico. Fra tutti i pianeti, Giove e Saturno sono quelli di massa maggiore, cosicché Newton ricercò perturbazioni orbitali nei loro moti. Avvalendosi della collaborazione dell'astronomo reale John Flamsteed, trovò che il moto orbitale di Saturno è perturbato quando i due pianeti si trovano alla loro distanza minima. Il processo di comparare ripetutamente la costruzione matematica con la realtà e successivamente modificarla in modo che fosse in accordo con essa condusse infine Newton a trattare i pianeti come corpi fisici, dotati di figura e volumi ben definiti.

Dopo avere modificato varie volte la sua costruzione matematica, Newton la applicò al sistema del mondo. Affermò che la forza di attrazione, che aveva derivato matematicamente, è la gravità universale. Trovò che la Luna si muove come se fosse attratta verso la Terra con una forza pari a 1/3600 della intensità della forza gravitazionale con cui la Terra attrae oggetti alla sua superficie. Poiché la Luna è 60 volte più lontana dal centro della Terra degli oggetti che si trovano alla superficie di questa, il fattore di 1/3600 è in accordo con la deduzione che la gravità della Terra si estende sino alla Luna e diminuisce proporzionalmente al quadrato della distanza.

La legge della gravitazionale universale spiega perché i pianeti seguano approssimativamente le leggi di Keplero e perché se ne allontanino nel modo in cui se ne allontanano, e dimostra perché (in assenza di attrito) tutti i corpi cadano con la stessa velocità in uno stesso luogo sulla Terra e perché la velocità di caduta vari con l'altezza e con la latitudine. La legge della gravitazione spiega anche i moti regolari e irregolari della Luna, fornisce una base fisica alla comprensione e alla previsione dei fenomeni di marea e dimostra che la precessione degli equinozi, che era stata osservata da molto tempo, ma non spiegata, è l'effetto dell'attrazione della Luna sul rigonfiamento equatoriale della Terra. Poiché la forza di attrazione matematica funzionava bene nello spiegare e nel prevedere i fenomeni osservati del mondo, Newton concluse che tale forza doveva «esistere veramente», anche se la filosofia tradizionale cui egli aderiva non consentiva e non poteva consentire che una tale forza fosse ammessa a far parte di un sistema della natura. Perciò egli si fece promotore di un'indagine sul modo in cui possano sorgere gli effetti della gravitazione universale.

Benché a volte Newton pensasse che la gravità universale potesse essere causata dagli impulsi di un flusso di particelle dell'etere che bombardassero un oggetto o da variazioni in un etere onnipervasivo, non suggerì nessuna di queste due nozioni nei *Principia* perché, come scrisse, non voleva introdurre ipotesi (*«hypotheses non fingo»*) come spiegazioni fisiche. Lo stile newtoniano lo aveva condotto a un concetto matematico della forza universale, e tale stile lo condusse ad applicare il suo risultato matematico al mondo fisico anche se non si trattava del tipo di forza in cui egli poteva credere.

Alcuni fra i contemporanei di Newton furono così turbati dall'idea di una forza di attrazione agente a distanza che non poterono indursi neppure a cominciare a indagarne le proprietà e trovarono difficile accettare la fisica newtoniana. Non poterono schierarsi al suo fianco, quando Newton disse che non era stato in grado di spiegare in che modo la gravità funzioni ma che «è sufficiente che la gravità esista veramente e basti a spiegare i fenomeni del cielo e delle maree». Coloro che accettavano lo stile newtoniano usarono la legge della gravitazione universale, dimostrarono come essa spiegasse molti altri fenomeni fisici e chiesero che si cercasse una spiegazione di come tale forza potes-

La velocità orbitale di un pianeta è inversamente proporzionale non alla distanza diretta fra il Sole e il pianeta, ma alla distanza perpendicolare (la linea tratteggiata fra il Sole e la tangente PP_1 all'orbita). Solo in due punti dell'orbita (il perielio e l'afelio) le due distanze coincidono.

$$\frac{v^2}{r} = \left(\frac{2\pi r}{T}\right)^2 \Big/ r = \left(\frac{4\pi^2 r^2}{T^2}\right)\frac{1}{r} =$$

$$4\pi^2\left(\frac{r^2}{T^2}\right)\frac{1}{r} = 4\pi^2\left(\frac{r^2 r}{T^2 r}\right)\frac{1}{r} =$$

$$4\pi^2\left(\frac{r^3}{T^2}\right)\frac{1}{r^2} = 4\pi^2 K\left(\frac{1}{r^2}\right)$$

La proporzionalità inversa della forza centripeta al quadrato della distanza per orbite circolari può essere ricavata deduttivamente a partire dalla terza legge kepleriana dei moti planetari e dalla legge della forza centripeta. Secondo la terza legge di Keplero, r^3/T^2 è una costante, K, dove r è il raggio dell'orbita del pianeta e T è il periodo dell'orbita. La legge della forza centripeta afferma che, per un'orbita circolare, la forza centripeta è v^2/r, dove v è la velocità del pianeta.

se essere trasmessa su enormi distanze attraverso uno spazio apparentemente vuoto. Lo stile newtoniano consentì a Newton di studiare la gravitazione universale senza premature inibizioni che avrebbero impedito la sua grande scoperta. Il biologo del Settecento Georges-Louis Leclerc de Buffon una volta ebbe occasione di scrivere che non si può distinguere lo stile di un uomo dall'uomo stesso. Nel caso di Newton, la sua massima scoperta non può essere separata dal suo stile.

La corrispondenza fra Hooke e Newton dimostra chiaramente che Hooke insegnò a Newton come analizzare il moto curvo. Hooke avanzò in seguito la rivendicazione molto più ambiziosa di aver suggerito a Newton la legge della gravitazione universale, secondo la quale la forza di gravità varia in proporzione inversa al quadrato della distanza. Molti storici hanno ripreso e avallato la tesi sostenuta da Hooke.

Questa rivendicazione però non regge. Hooke aveva suggerito semplicemente che i pianeti sono soggetti a una forza che varia in proporzione inversa al quadrato della distanza e che è diretta verso il Sole. La gravitazione universale è molto di più di una forza diretta verso il Sole. Essa implica anche un effetto dei pianeti sul Sole. Inoltre, essa si applica a tutti gli oggetti dell'universo. La legge della gravitazione universale non è semplicemente una relazione che esprime una proporzione inversa a un quadrato; essa è anche una relazione matematica fra le masse dei corpi che si attraggono fra loro. Ci volle una capacità di penetrazione grandissima per passare da una forza diretta verso il Sole che obbediva a una proporzionalità inversa al quadrato della distanza alla gravitazione universale. E ci volle il genio di Newton per inventare il concetto moderno di massa.

Newton non ritenne di dovere qualcosa a Hooke neppure per il suggerimento che la forza centripeta sia inversamente proporzionale al quadrato della distanza. Nel 1673 Huygens aveva pubblicato un supplemento a un libro sull'orologio a pendolo in cui si afferma che nel moto circolare la forza centrifuga è misurata da v^2/r, dove v è la velocità del corpo orbitante e r è il raggio di rotazione. Newton aveva scoperto indipendentemente la stessa relazione fra il 1660 e il 1670. Poiché la differenza matematica fra una forza centrifuga e una forza centripeta è solo una questione di direzione, la relazione v^2/r vale anche per la forza centripeta. Da questa relazione e dalla terza legge di Keplero segue, mediante operazioni algebriche semplici, che la forza centripeta varia in proporzione inversa al quadrato della distanza.

Dopo la pubblicazione del libro di Huygens chiunque avesse avuto una conoscenza rudimentale dell'algebra avrebbe potuto trovare una forza centripeta variabile in proporzione inversa al quadrato del raggio per un'orbita circolare. Perciò Newton non ritenne necessario attribuire a Hooke il merito di aver formulato una legge di proporzionalità inversa al quadrato della distanza.

Tanto Hooke quanto Newton erano ben consapevoli che trovare una legge di proporzionalità inversa al quadrato della distanza per orbite circolari non era la stessa cosa che dimostrare che la legge valeva per orbite ellittiche in cui il moto obbedisce alla legge delle aree di Keplero. Il compito di Newton consistette nel dimostrare che una legge della proporzionalità inversa della forza centripeta al quadrato della distanza corrispondeva al moto orbitale secondo le leggi kepleriane delle orbite ellittiche e delle aree. Esaminando questo punto nella sua lettera datata 6 gennaio 1680, Hooke commise un errore fondamentale che deve aver convinto Newton del fatto che Hooke non capiva completamente ciò di cui stava parlando. Hooke scrisse che, se l'attrazione varia in proporzione inversa al quadrato della distanza, la velocità orbitale di un pianeta sarà, «come suppone Keplero, reciproca alla distanza». Nelle condizioni supposte da Hooke, la velocità orbitale non è però inversamente proporzionale alla distanza diretta dal Sole tranne che nei punti estremi dell'orbita: il perielio e l'afelio. In considerazione di questo errore di Hooke, Newton non era disposto a riconoscergli il merito di aver suggerito la proporzionalità inversa della forza centripeta al quadrato della distanza.

Nel 1717 Newton, desiderando assicurarsi la priorità nella scoperta della legge della gravitazione, con la sua proporzionalità inversa al quadrato della distanza, si inventò la storia secondo cui si sarebbe occupato del moto della Luna come di un caso di caduta libera non mentre stava scrivendo i *Principia*, ma due decenni prima, negli anni sessanta. I documenti di quel periodo indicano però che egli allora non stava confrontando la caduta della Luna nella sua orbita con la caduta di oggetti negli ambienti terrestri, ma stava confrontando la «tendenza centrifuga» della Luna nella sua orbita con la «tendenza centrifuga» di un corpo sulla superficie della Terra, partecipe della rotazione diurna della Terra stessa. Newton calcolò che per orbite planetarie circolari la «tendenza centrifuga» doveva essere inversamente proporzionale alla distanza del pianeta dal Sole, ma non trasse dal calcolo alcuna conclusione fisica.

Newton non pubblicò mai questo falso resoconto del suo interesse per il moto della Luna. Lo incluse nella minuta manoscritta di una lettera al francese Pierre Des Maizeaux, ma poi lo cancellò. Newton fece circolare anche l'aneddoto familiare della caduta di una mela dall'albero che avrebbe dato l'avvio alla catena di riflessioni che lo condussero poi alla scoperta della gravitazione universale. Presumibilmente anche quest'invenzione rientra nella sua campagna per far risalire la scoperta della gravità, o almeno le radici della scoperta, a una ventina d'anni prima dei *Principia*.

Le vere radici della scoperta non possono essere datate a prima del dicembre 1684, quando Newton riconobbe per la prima volta che, se il Sole attrae la Terra, la Terra deve attrarre il Sole con una forza di ugual grandezza. Nel 1685 egli superò la sua abituale riluttanza a scrivere per esteso le sue scoperte e cominciò a redigere i *Principia* in vista della loro pubblicazione per opera della Royal Society. Forse la sua disponibilità a presentare la sua opera all'esame pubblico (e quindi al rischio di una possibile disapprovazione) fu motivata in un primo tempo dall'importanza della sua scoperta delle perturbazioni planetarie, seguita dalla sua audace concezione della gravitazione universale. Era ora alla sua portata la fondazione di un nuovo sistema di filosofia naturale che poteva essere esposto sulla base di principi matematici. In breve, una volta che Newton ebbe veramente qualcosa di importante da dire sulla dinamica celeste, fu disposto e anche impaziente di presentarla al mondo.

TEMI METAMAGICI

di Douglas R. Hofstadter

I cubisti giocherellano con i cubetti del Cubo Magico e i cubologhi li risolvono

Cubitis magikia, *n.* Grave disturbo mentale accompagnato da un prurito alle dita alleviabile solo mediante un prolungato contatto con un cubo multicolore di origine ungherese e giapponese. I sintomi durano spesso diversi mesi. Altamente contagioso.

Ciò che questa asettica voce di dizionario medico non dice è che il contatto con il cubo multicolore non solo cura il prurito, ma lo provoca anche. Inoltre, essa trascura di rilevare che il disturbo può essere molto piacevole. Lo posso ben dire io che ne ho sofferto l'anno scorso e ancora ne ho i sintomi.

Bűvös Kocka - il Cubo Magico, noto anche come cubo di Rubik - ha messo simultaneamente a soqquadro il mondo dei rompicapo, della matematica e dei calcolatori. Raramente un rompicapo ha acceso a tal punto l'immaginazione della gente. Per trovare un precedente analogo bisogna riandare al famoso gioco del 15 di Sam Loyd, che provocò nel XIX secolo una vera follia collettiva e rimane uno dei più popolari rompicapo del mondo. Il gioco del 15 e il Cubo Magico sono spiritualmente affini: il primo è un problema bidimensionale che richiede di riportare nella loro giusta posizione i pezzi numerati e mischiati di un quadrato 4×4, il secondo è un problema tridimensionale che richiede di riportare nella loro giusta posizione i pezzi colorati e mischiati di un cubo 3×3×3. Le soluzioni di entrambi i problemi richiedono la ripetuta distruzione dei preziosi progressi raggiunti; non c'è modo di raggiungere l'obiettivo senza una parziale, ma temporanea distruzione dell'ordine conseguito fino a quel punto. Se ciò è difficile da imparare con il gioco del 15, ancora di più lo è con il Cubo Magico. Entrambi i rompicapo hanno inoltre la diabolica proprietà che se qualcuno li smonta e li rimonta in certe posizioni apparentemente innocue, il risultato diventa assolutamente irraggiungibile, con conseguenti indescrivibili angosce del potenziale solutore.

Il Cubo Magico è molto più di un semplice rompicapo: è un'ingegnosa invenzione meccanica, un passatempo, uno strumento didattico, una fonte di metafore, un'ispirazione. Adesso sembra ormai un oggetto inevitabile, ma la sua scoperta richiese molto tempo. I tempi erano però maturi, perché l'idea germinò e si sviluppò quasi parallelamente in Ungheria e in Giappone e forse anche altrove. Recentemente un funzionario del governo francese ha affermato di aver visto un cubo del genere, fatto in legno, a Istanbul nel 1920 e poi ancora a Marsiglia nel 1935. Mancano però conferme e l'affermazione sembra dubbia. Resta comunque che il lavoro di Rubik venne completato nel corso del 1975 e il suo brevetto ungherese porta quella data. In modo del tutto indipendente, Terutoshi Ishige, un ingegnere autodidatta proprietario di una piccola acciaieria vicino a Tokio, arrivò meno di un anno dopo a un progetto del tutto simile a quello di Rubik e ottenne un brevetto giapponese nel 1976. Anche a Ishige va il merito di questa meravigliosa intuizione.

Chi è Rubik? Ernő Rubik insegna architettura e disegno presso la scuola per artisti di Budapest. Cercando di affinare la capacità dei suoi studenti di visualizzare oggetti tridimensionali, ebbe l'idea di un cubo 3×3×3 le cui sei facce 3×3 potessero ruotare intorno al centro, ma in modo che l'intero cubo non si disfacesse. Ogni faccia veniva inizialmente colorata in modo uniforme, ma ripetute rotazioni delle varie facce mescolavano spaventosamente i colori. Gli studenti dovevano poi trovare il modo per ripristinare l'ordine iniziale.

Quando mi descrissero per la prima volta il cubo al telefono, mi sembrò impossibile da un punto di vista fisico. Risulta logico pensare che esso si disfi nei suoi «cubetti» costituenti. A che cosa rimane attaccato un cubetto d'angolo? Se si immagina di ruotare ognuna delle tre facce a cui appartiene, ci si rende conto che il cubetto d'angolo in questione si può staccare da ciascuno dei tre cubetti di spigolo confinanti. Allora cosa lo tiene a posto? Alcuni pensano subito a calamite, elastici o elaborati sistemi di cavetti intrecciati all'interno del cubo. Il progetto è invece molto semplice e non ricorre a espedienti di questo tipo.

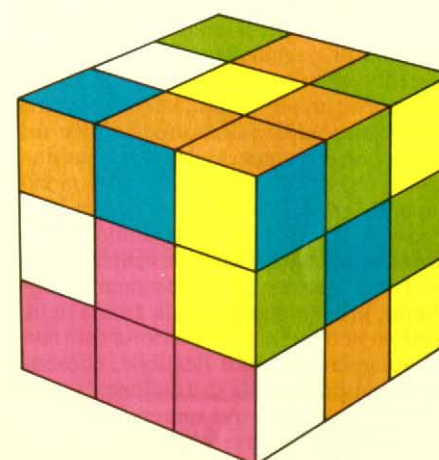
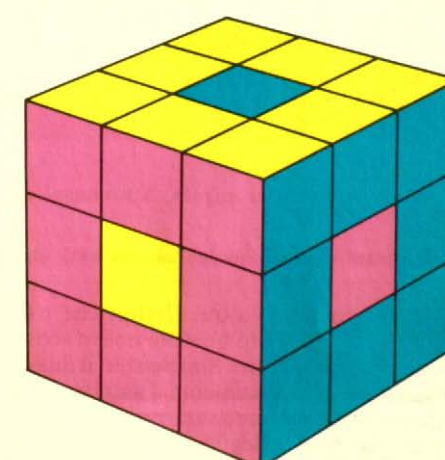
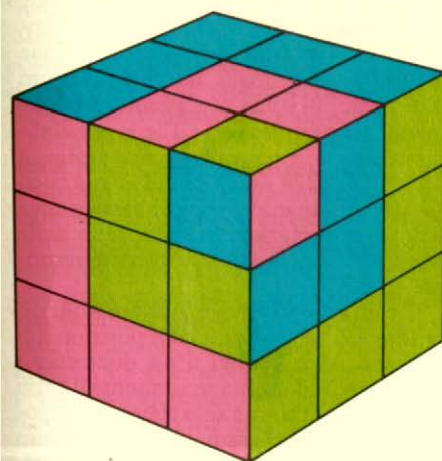
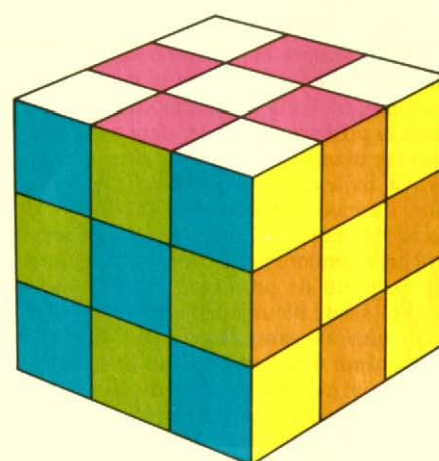
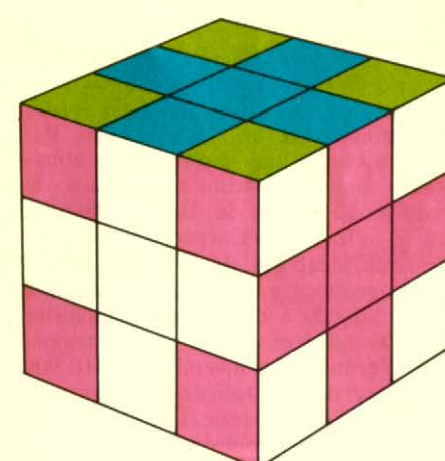
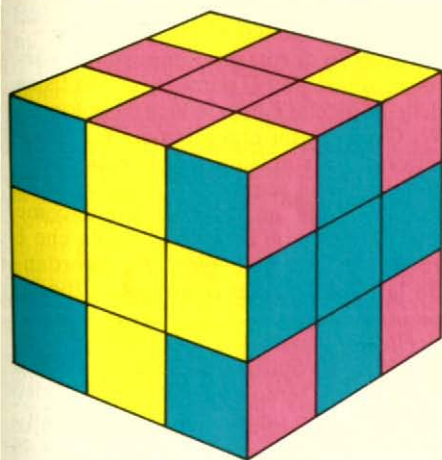
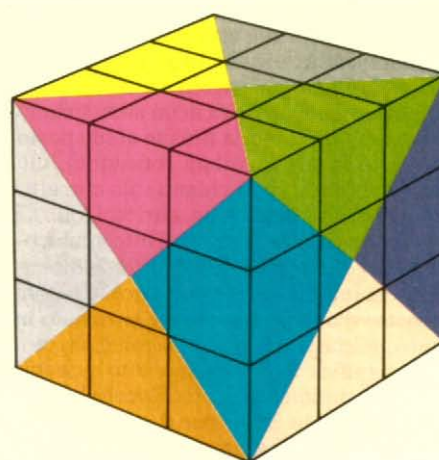
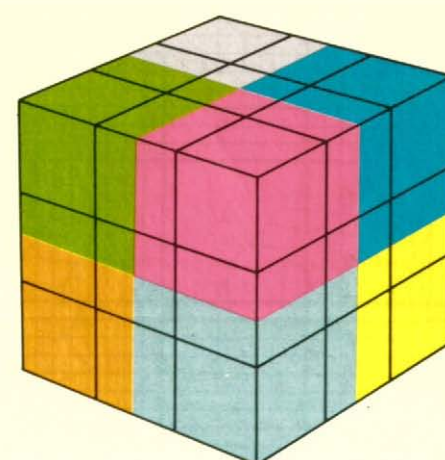
In realtà, il Cubo Magico può essere smontato in pochi secondi *(si veda la figura in basso a pagina 141)*, rivelando una struttura interna così semplice che è difficile capire come possa funzionare. Il cubo si disfa davvero. Per capire cosa lo tenga insieme si osservi per prima cosa che ci sono tre tipi di cubetti: sei centrali, dodici di spigolo e otto d'angolo. I cubetti centrali hanno solo una «faccetta», i cubetti di spigolo ne hanno due e i cubetti d'angolo ne hanno tre. Per di più i sei cubetti centrali non sono in realtà del tutto cubici - sono semplici facciate attaccate ad assi che nascono da un perno a sei bracci posto al centro. Gli altri cubetti sono dei cubi quasi completi, tranne per il fatto che hanno tutti un piccolo «piede» smussato che si protende verso il centro del cubo e degli intagli curvi rivolti verso l'interno.

Il trucco sta essenzialmente nel fatto che i cubetti si sostengono l'un l'altro mediante i loro piedi, senza che nessun cubetto sia attaccato a un altro. I cubetti di spigolo trattengono i piedi dei cubetti d'angolo, i cubetti d'angolo trattengono i piedi dei cubetti di spigolo. I cubetti centrali sono le chiavi di volta. Quando uno strato, per esempio quello superiore, compie una rotazione, mantiene esso stesso la sua compattezza orizzontalmente ed è tenuto in sede verticalmente dal suo stesso centro e dallo strato equatoriale al di sotto. Lo strato equatoriale ha una traccia circolare cava (formata dagli intagli nei suoi cubetti) che guida il movimento dei piedi dello strato superiore e aiuta a tenere insieme lo strato superiore.

Nel suo fondamentale trattato *Notes on Rubik's «Magic Cube»*, David Singmaster, professore di scienze matematiche e di macchine calcolatrici al Politecnico della South Bank di Londra, definisce «problema meccanico base» l'immaginare come sia costruito il cubo. Mi chiedo a volte se il compito di visualizzazione di cui parlava Rubik per i suoi studenti consisteva nel risolvere il problema del ripristino dell'ordine (chiamato da Singmaster il «problema matematico base») o nel risolvere il problema meccanico. Io sospetto che quest'ultimo sia il più difficile dei due. Io stesso ho impiegato più di 50 ore di lavoro, distribuite su più mesi, per risolvere il problema del ripristino dell'ordine, ma non sono riuscito a risolvere il problema meccanico prima di vedere il cubo smontato. Secondo una stima informale di Singmaster, le persone che arrivano a risolvere il problema del ripristino dell'ordine (senza suggerimenti) impiegano in media due settimane di sforzi concentrati. Naturalmente, è difficile per chiunque sia riuscito a risolvere il problema dire esattamente quanto ci ha messo (come si può distinguere il gioco dal lavoro?), ma si può tranquillamente sostenere che se siete destinati a risolvere interamente il problema del ripristino dell'ordine, impiegherete un tempo compreso tra le cinque ore e un anno. Penso che la cosa sia rassicurante.

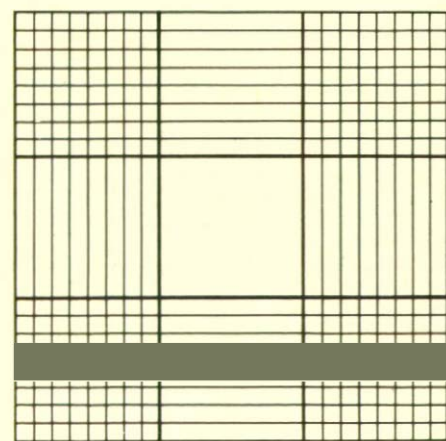
Un fatto importante di cui molti non tengono conto fin dall'inizio è che riportare il cubo alla posizione di partenza (lo stadio in cui ogni faccia è di colore uniforme) è così difficile da rendere necessario trovare un algoritmo generale per arrivare al risultato a partire da un qualsiasi stadio di mescolamento. Nessuno può riportare un Cubo Magico mescolato al suo stadio iniziale procedendo semplicemente per tentativi ed errori. Chiunque riesca a ritornare alla posizione di partenza ha costruito una piccola scienza.

Una parola d'avvertimento: le soluzioni proposte per il problema meccanico mancano spesso di chiarezza e hanno o troppi dettagli o troppo pochi. È certamente un'impresa venire a capo di un meccanismo con le multiformi capacità di contorsione del Cubo Magico, ma è forse un'impresa non da meno descrivere il meccanismo con un linguaggio e dei diagrammi che risultino facilmente comprensibili. Allo stesso modo, è necessario adottare una chiara notazione per rendere conto degli algoritmi che ripristinano lo stadio iniziale del cubo. Lo stesso Singmaster utilizza un'ottima notazione che è ormai entrata nell'uso comune e di cui parlerò più avanti. Un'altra parola di avvertimento: io non sono un «cubologo» (inteso come uno che ha dato un contributo alla profonda scienza della cubologia). Sono un semplice cubista, un dilettante affascinato dal cubo e dai virtuosi che ne hanno approfondito la conoscenza. Non sono quindi la persona adatta a ricevere nuove soluzioni del problema meccanico o del problema del ripristino dell'ordine, soluzioni che ammontano ormai a centi-

Il Cubo Magico nella posizione di partenza.

Un cubo mescolato.

La «bella disposizione» detta Punti.

Il Mesone Gigante con la posizione Ciliegine.

La configurazione Pons Asinorum.

La configurazione Croce di Natale.

La configurazione Croce di Plummer.

Una colorazione alternata con 12 colori.

Una colorazione alternata con otto colori.

F *sta per faccia,* **S** *per spigolo,* **A** *per angolo.*

naia. Raccomando a quei lettori che ritengono di aver fatto qualche nuova scoperta di comunicarlo a Singmaster, il quale le aggiorna periodicamente il suo lavoro. Il libro di Singmaster sarà presto distribuito alle librerie degli Stati Uniti dall'editore Enslow di Hillside, N. J., a un prezzo di 5,95 dollari. L'indirizzo di Singmaster è: Department of Mathematical Sciences and Computing, Polytechnic of the South Bank, London SEI OAA, Inghilterra.

Sono certo che ormai l'interesse del lettore sia stato solleticato al punto da fargli ritenere urgente e prioritario possedere immediatamente un Cubo Magico. Per fortuna la cosa può essere risolta abbastanza facilmente: cubi magici sono in produzione presso la Logical Games, Inc., e la Ideal Toy Corporation. Il nome versione Ideal, ma entrambe le versioni sono ugualmente «magiche». In Italia il Cubo Magico è distribuito in esclusiva dalla Mondadori Giochi, Sommacampagna (Verona) ed è in vendita a L. 10 900 nei negozi di giocattoli, nei grandi magazzini e nelle cartolibrerie. È probabile che molte persone li compreranno senza sospettare la profonda difficoltà del «problema matematico base». Gireranno innocentemente quattro o cinque facce e si troveranno improvvisamente perduti e senza speranza. Poi inizieranno, magari freneticamente, a girare faccia dopo faccia in un modo e in un altro, accorgendosi a poco a poco di aver perso irrimediabilmente qualcosa di prezioso. Si sentiranno come un bambino che vede un palloncino salire verso il cielo.

È assodato che un cubo può essere randomizzato con pochi giri: serva d'avvertimento per il principiante. Molti cercano all'inizio di riagguantare la posizione di partenza sistemando dapprima una sola faccia. Poi, un po' in crisi, lasciano in giro il cubo parzialmente risolto, là dove un amico può individuarlo. La ben nota sindrome del «Non toccare» entra in gioco quando l'amico prende in mano il cubo e dice «Cos'è?». Il potenziale solutore, terrorizzato dall'idea che possa venir distrutto il risultato tanto faticosamente raggiunto, urla «Non toccarlo!» Ironia della sorte, si può arrivare alla vittoria solo con un atteggiamento più flessibile, consentendo proprio quella distruzione.

Il principiante prova un terribile senso di irreversibilità davanti alla distruzione della posizione di partenza, una specie di paura di cadere in un precipizio. Quando il mio primo cubo (ora ne ho cinque) venne messo sottosopra (da un ospite), mi sentii a un tempo sollevato (perché era inevitabile) e triste (perché temevo perduta per sempre la posizione iniziale). Al fisico che c'è in me venne in mente l'entropia. Una volta irrimediabilmente perduta la posizione iniziale, ogni nuova rotazione di una faccia o di un'altra mi sembrava irrilevante. Nella mia ingenuità, non riuscivo a distinguere uno stadio di mescolamento da un altro, così come apparentemente nulla distingue un piatto di spaghetti da un altro, un mucchio di foglie cadute da un altro e così via. I dettagli non avevano alcun significato per me, quindi non li registravo. Successivamente nei miei lunghi tentativi di soluzione «a caso» mi apparve vivida e chiara la vastità dello spazio di possibili mescolamenti dei cubetti.

Come nel caso di un mazzo di carte, si può calcolare il numero esatto di possibili ricombinazioni del cubo. Una stima iniziale potrebbe essere eseguita nel modo seguente. La prima osservazione - piuttosto elementare - è che ruotando una faccia qualsiasi ogni angolo va in un altro angolo, ogni spigolo va in un altro spigolo e il centro della faccia rimane al suo posto (a parte la sua invisibile rotazione). Gli angoli, quindi, si mischiano solo con altri angoli e gli spigoli con altri spigoli. Ci sono otto cubetti d'angolo e otto «cubicoli» d'angolo (le nicchie di spazio, indipendentemente dal loro contenuto). Ciascun cubetto d'angolo può essere sistemato in uno qualsiasi degli otto cubicoli d'angolo. Ciò significa che abbiamo otto possibilità di riempimento per il cubicolo 1, sette per il cubicolo 2, sei per il cubicolo 3, e così via. Quindi gli angoli possono essere sistemati nei loro cubicoli in $8\times7\times6\times5\times4\times3\times2\times1$ (= 8!) differenti modi. Ogni angolo, però, può essere in uno qualsiasi dei tre orientamenti. Ci si aspetterebbe quindi un ulteriore fattore di 3^8 dagli angoli. Lo stesso ci si aspetterebbe per i cubetti di spigolo: 12 oggetti possono essere permutati tra di loro in 12! differenti modi e se poi ciascuno di essi ha due possibili orientamenti si arriva a un altro fattore di 2^{12}. I cubetti di centro non lasciano mai le loro posizioni di partenza (a meno che non si ruoti l'intero cubo), quindi non contribuiscono. Se moltiplichiamo i numeri ottenuti, arriviamo a 519 024 039 293 878 272 000 possibili posizioni, circa 5×10^{20}.

Ma sotto c'è un'assunzione: che qualsiasi cubetto possa andare in qualsiasi cubicolo in qualsiasi orientamento, indipendentemente dalla posizione e dall'orientamento degli altri cubetti. Come vedremo, questo non è il caso. Risulta esserci una limitazione all'orientamento dei cubetti d'angolo: sette qualsiasi di essi possono essere orientati in modo arbitrario, ma l'ultimo ha allora un orientamento forzato, togliendo così un fattore pari a tre. Un'analoga limitazione vale per i cubetti di spigolo: dei 12, 11 possono essere orientati arbitrariamente, ma l'ultimo risulta determinato, eliminando così un altro fattore pari a due. C'è un'ultima limitazione sulle permutazioni dei cubetti (indipendentemente dal loro orientamento), secondo la quale tutti i cubetti tranne due possono essere sistemati dove si vuole, ma gli ultimi due hanno una posizione forzata. Questo toglie un ultimo fattore pari a due; la stima precedente si riduce quindi di un fattore pari a 12 e le possibilità scendono a un valore pari «solo» a 43 252 003 274 489 856 000, ossia circa 4×10^{19}. Si supera ancora abbastanza, va detto, l'affermazione sull'etichetta della Ideal: «Più di tre miliardi di combinazioni».

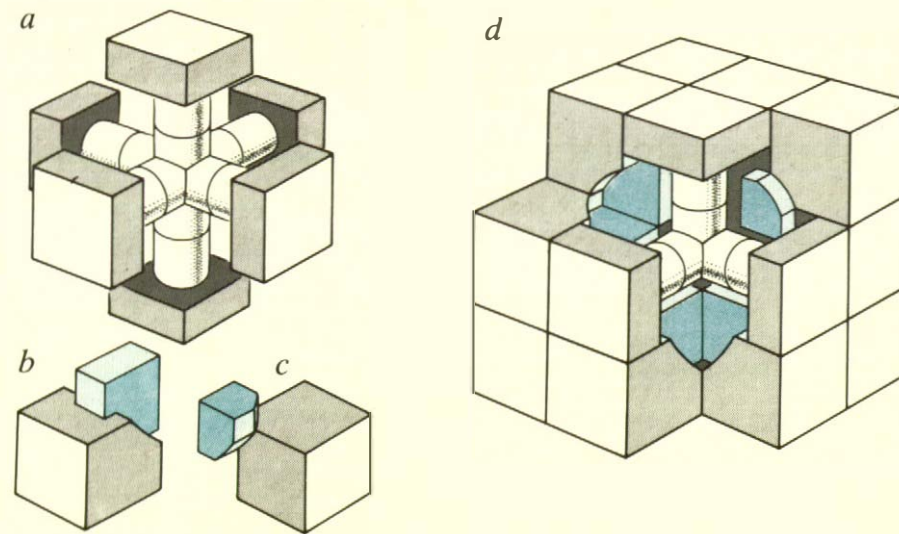
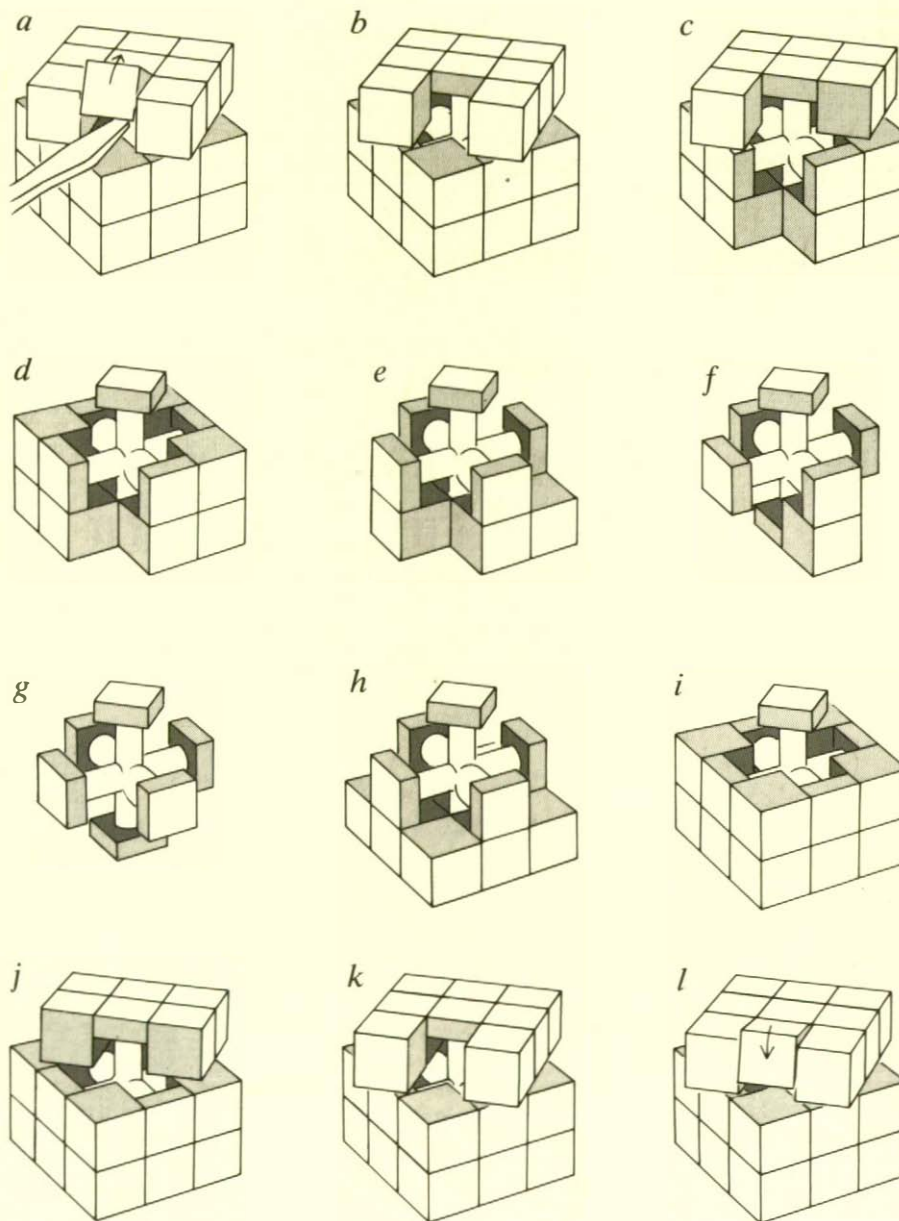
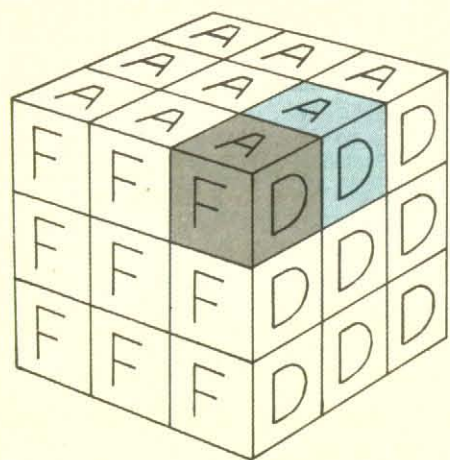
C'è un altro modo per arrivare a questo fattore 12: se si inizia dalla posizione di partenza, si è limitati a un dodicesimo degli «ovvi» stadi, ma se si smonta il cubo e lo si rimonta con un solo cubetto d'angolo ruotato di 120 gradi, ci si trova in uno stadio precedentemente irraggiungibile, da cui è ora raggiungibile una famiglia intera di 43 252 003 274 489 856 000 nuovi stadi. Ci sono 12 simili famiglie non coincidenti di stati del cubo, abitualmente chiamate orbite dai teorici dei gruppi.

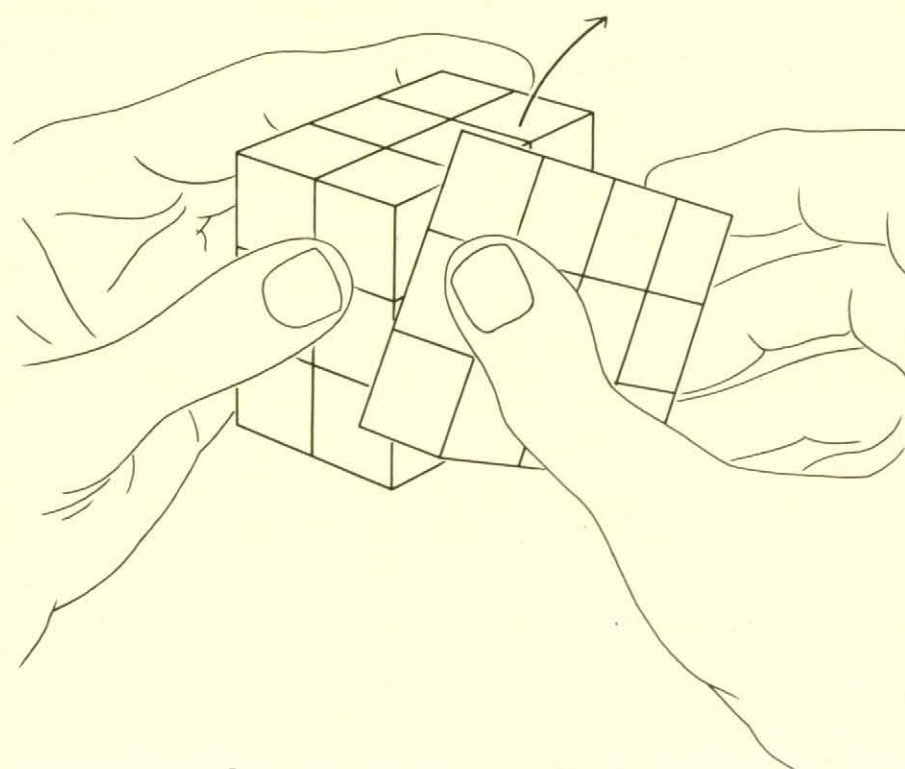
Parlando di rotazioni impossibili, vorrei ricordare un bel parallelo con la fisica delle particelle messo in luce da Solomon W. Golomb (un nome familiare a molti lettori di Martin Gardner). È impossibile compiere una sequenza di movimenti al termine della quale tutto risulti come prima, tranne un cubetto d'angolo che è ruotato di un terzo di giro. Ora, ricordando la famosa ipotetica particella fondamentale con carica +1/3 e la sua antiparticella con carica − 1/3, Golomb chiama un terzo di rotazione in senso orario un quark e un terzo di rotazione in senso antiorario un antiquark. Come i loro omonimi cubici, le particelle quark si sono dimostrate provocantemente elusive e molti fisici teorici credono ora nel confinamento del quark: il principio secondo cui è impossibile avere un quark (o antiquark) libero isolato. Questa correlazione tra quark cubici e particelle quark è veramente bella.

In realtà, la connessione è ancora più profonda. Non possono esistere particelle quark libere, ma possono esistere legate insieme in gruppi: una coppia quark-antiquark è un mesone e un trio di quark con carica intera è un barione. (Ne è un esempio il protone, con carica +1.) Ora, cosa abbastanza sorprendente, nel Cubo Magico è possibile dare a due cubetti d'angolo rotazioni di un terzo, sempre che siano in direzioni *opposte* (una in senso orario, l'altra in senso antiorario). È anche possibile dare a tre cubetti d'angolo rotazioni di un terzo, sempre che siano nella *stessa* direzione. Così Golomb chiama uno stato con due angoli ruotati in senso opposto un mesone e uno con tre angoli ruotati nella stessa direzione un barione. Nel mondo delle particelle, possono esistere solo combinazioni di quark con carica complessiva intera. Nel mondo del cubo sono consentite solo combinazioni di quark con una somma intera di rotazioni. È questo un altro modo per dire che l'orientamento dell'ottavo cubetto d'angolo è sempre forzato dall'orientamento dei primi sette. Nel mondo del cubo, la ragione del confinamento dei quark va ricercata nella teoria dei gruppi. Esiste forse, nella teoria dei gruppi, una spiegazione strettamente connessa con la prima per il confinamento delle particelle quark. Questo rimane da vedere, ma in ogni caso il parallelo è stimolante e divertente.

Se abbiamo un cubo primitivo (un cubo nella posizione di partenza), che tipo di sequenza di mosse creerà un mesone o un barione? Abbiamo qui un esempio dell'idea più potente in cubologia: l'idea di sequenze di movimenti standard che raggiungono particolari riordinamenti di alcuni cubi, lasciando tutto il resto intoccato (invarianti, come dicono i teorici dei gruppi). Ci sono molti termini differenti per queste sequenze di movimenti standard. Le ho sentite chiamare operatori, trasformazioni, parole, strumenti, processi, manovre, routine, subroutine e maero; i primi tre termini appartengono alla teoria dei gruppi e gli ultimi tre sono adattati dal gergo della scienza dei calcolatori. Ciascun termine ha il suo pregio e mi piace usarli tutti.

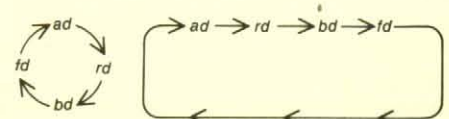
Per parlare dei processi ci vuole precisione e ciò significa avere una buona notazione tecnica. Presenterò pertanto la notazione di Singmaster. Abbiamo innanzitutto bisogno di un modo per riferirci a una qualsiasi particolare faccia del cubo. Una possibilità consiste nell'usare i nomi dei colori come nomi delle facce, anche dopo che i colori dei cubetti si sono mescolati. Potrebbe sembrare insensato chiamare bianca una faccia se il bianco è sparpagliato dappertutto. Si ricordi, però, che il cubetto centrale bianco non si muove mai rispetto agli altri cinque cubetti centrali e definisce quindi la «faccia base» per il bianco. Allora perché non usare i nomi dei colori per le facce? Il problema

Disposizione di una qualsiasi faccia del Gruppo fetta.

Perno **(a)**, *cubetto di spigolo* **(b)**, *cubetto d'angolo* **(c)** *e visione del meccanismo interno* **(d)**.

Smontaggio **(a-f)** *e montaggio* **(g-l)**. *È sempre possibile ristabilire la posizione di partenza!*

Il cubetto **adf** *(grigio) e il cubetto* **ad** *(colorato).*

sta nel fatto che i vari cubi differiscono tra di loro per la disposizione dei colori: persino due cubi di uno stesso fabbricante possono avere posizioni di partenza diverse. Una convenzione più generale consiste nel chiamare le facce semplicemente «sinistra» e «destra», «davanti» e «dietro», «sopra» e «sotto». Per evitare ripetizioni di iniziali si può sostituire le parole «davanti» e «dietro» con le parole «fronte» e «retro», e le parole «sopra» e «sotto» con le parole «alto» e «basso». Abbiamo ora sei facce: *S, D, F, R, A, B.* Qualsiasi particolare cubetto può essere designato utilizzando le lettere minuscole delle facce a cui esso appartiene. Così *ad* (o *da*) sta per il cubetto di spigolo sul lato destro dello strato superiore e *adf* per il cubetto d'angolo di fronte ad esso *(si veda la figura in alto).*

La mossa più spontanea per un cubista che usi la destra sembra essere quella che consiste nel prendere la faccia di destra con il pollice puntato sulla faccia frontale e muovere in avanti il pollice. Vista dal lato destro, questa operazione provoca un quarto di rotazione oraria della faccia *D.* Questo movimento sarà designato con la lettera *D (si veda la figura qui sotto).* La mossa speculare alla precedente, in cui la mano sinistra gira il lato *S* in senso antiorario (se visto da sinistra), è S^{-1}, o più brevemente *S'*. Una rotazione oraria del lato *S* è chiamata, naturalmente, *S.* Una rotazione oraria di 90 gradi di una faccia (dal punto di vista di un osservatore che guardi il centro di quella faccia) prende il nome della lettera di quella faccia, e il suo inverso - il quarto di rotazione in senso antiorario - ha un apice (') o un esponente -1 che segue l'iniziale della faccia. Le rotazioni di un quarto di giro saranno d'ora in avanti chiamate rotazioni *q.*

Con questa nomenclatura possiamo ora scrivere una qualsiasi sequenza di mosse, indipendentemente dalla sua complessità. Un esempio banale è dato da quattro successivi *D*, che denoteremo con D^4. Nel linguaggio della teoria dei gruppi questa è l'operazione identità: il suo effetto è nullo. Tale sequenza può essere espressa con l'equazione $R^4 = I$, dove qui la *I* sta per l'«azione» di non far nulla.

Supponiamo di ruotare due facce differenti, per esempio prima *D* poi *A*. Questa azione sarà denotata con *DA*, non con *AD*. Si noti prima di tutto che *DA* e *AD* hanno effetti ben differenti. Per verificarlo, si effettui dapprima *DA* su un cubo primitivo, si notino i suoi effetti e poi si torni indietro, si provi con *AD* e si veda come sono differenti i suoi effetti. L'inverso di *DA* è ovviamente *A'D'*, non *D'A'*. (Detto per inciso, questa strategia consistente nell'esperimentare sequenze di mosse su un cubo primitivo è di grande utilità. Io mi accorsi molto presto di quanto fosse utile comprare un secondo cubo per poter lavorare alla soluzione su uno e fare esperimenti sull'altro, non facendo mai allontanare il secondo dalla posizione di partenza.)

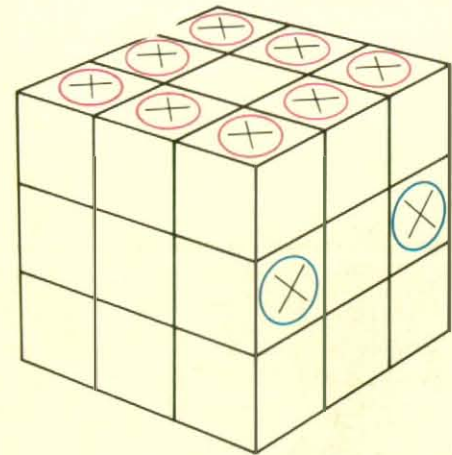
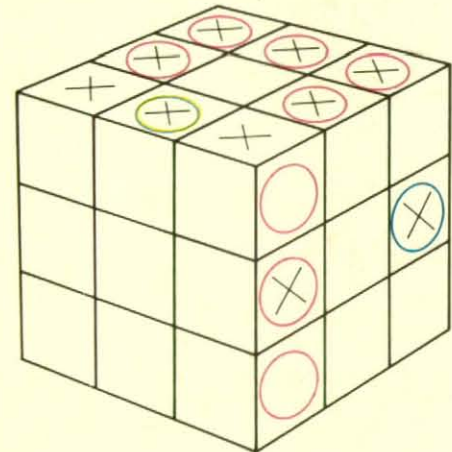
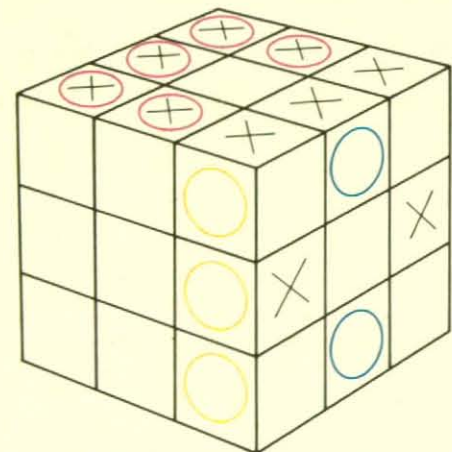
Qual è l'effetto di una certa «parola»? Vale a dire, quali cubetti si muovono e dove vanno? Per rispondere alla domanda abbiamo bisogno di una notazione per i movimenti dei singoli cubetti. La mossa *D* ha l'effetto di portare il cubetto *ad* sulla faccia posteriore a occupare il cubicolo *rd.* Allo stesso tempo, il cubetto *rd* passa sotto a occupare la posizione *bd*, il cubetto *bd* si sposta per occupare il cubicolo *fd* e il cubetto *fd* sale in alto in *ad.* Simbolicamente, tutto ciò si può rappresentare con l'uno o l'altro dei seguenti disegni:

Questo si chiama un ciclo-4 e lo scriveremo in modo più compatto: *(ad, rd, bd, fd).* Naturalmente, non ha importanza il punto da cui partiamo; potremmo benissimo scrivere anche *(rd, bd, fd, ad).* Ha invece importanza l'ordine delle lettere nel nome dei cubetti. Possiamo rovesciarle tutte o nessuna, ma non alcune. Se si immagina che le lettere designino le faccette, la cosa risulterà chiara. Per esempio, se avessimo scritto *(ad, dr, bd, df)*, avremmo rappresentato un ciclo-4 con in gioco gli stessi cubicoli, ma in cui ciascun cubetto si è mosso prima di spostarsi da un cubicolo al successivo. Naturalmente, un ciclo di questo tipo non può essere compiuto con una sola rotazione *q*, ma deve essere il risultato di una *sequenza* di rotazioni *q* di differenti facce (un operatore). Oppure si consideri il seguente ciclo-8: *(ad, af, as, ar, da, fa, sa, ra).* Questo ciclo ha lunghezza otto, ma mette in gioco solo quattro cubicoli. Ogni cubetto, dopo aver fatto un intero giro intorno alla faccia superiore, torna indietro rovesciato. Dopo due giri completi torna com'era partito. Ogni faccetta ha fatto una «striscia di Möbius». Possiamo designare questo «ciclo-4 rovesciato» con $(ad, af, as, ar)_+$, dove il segno «più» indica il rovesciamento. Andrebbero bene anche la denotazione $(da, fa, sa, ra)_+$ e numerose altre. Così la notazione di ciclo ci dice non solo dove va un cubetto, ma anche il suo orientamento rispetto agli altri cubetti nel suo ciclo.

Per completare la nostra descrizione dell'effetto di *D*, dobbiamo trovare una notazione per il ciclo-4 degli angoli. Come nel caso degli spigoli, siamo liberi di partire dall'angolo che vogliamo e dobbiamo stare attenti a seguire le faccette in modo da avere i giusti orientamenti. Inoltre, *D* ha un effetto piuttosto banale sugli angoli: *(adf, rda, bdr, fdb)*, che si potrebbe anche scrivere *(dar, drb, dbf,*

D, ***una mossa spontanea per un cubista destrorso.***

dfa) e in molti altri modi. Riassumendo, possiamo scrivere *D = (ad, rd, bd, fd)(adf, rda, bdr, fdb)*. Questo ci dice che *D* consiste di due cicli-4 disgiunti. (Se volessimo, potremmo trovare un termine per la rotazione di 90 gradi del cubetto centrale della faccia *D*, ma dato che questa rotazione è invisibile non lo faremo.)

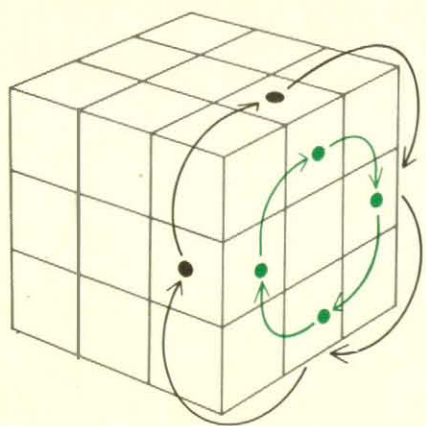
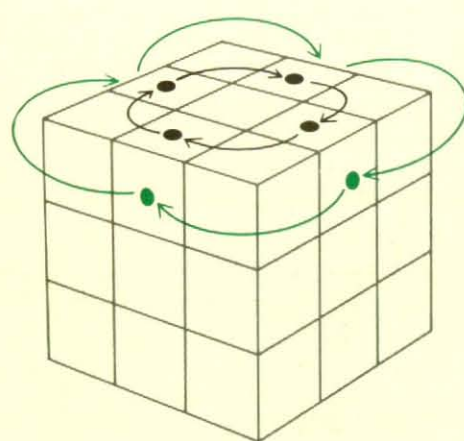
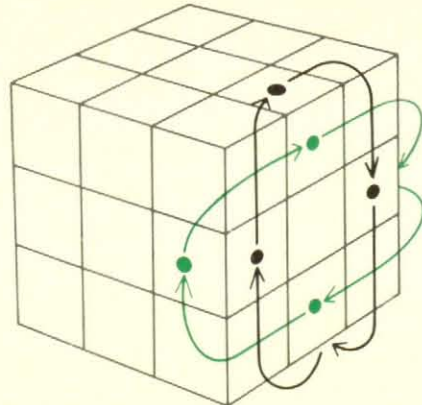
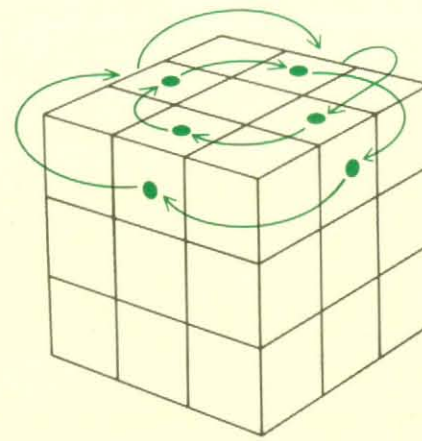
E per denotare una sequenza di mosse come *DA?* Bene, prendiamo un cubo primitivo ed effettuiamo *DA*. Partiamo poi da un cubetto qualsiasi che si sia mosso e descriviamo la sua traiettoria. Per esempio, *ad* è andato in *rd*. Perciò *rd* si è spostato; dove è andato? Troviamo la nuova collocazione di quel cubetto (è *bd*) e continuiamo a inseguire i cubetti intorno al cubo fino a trovare quello che si è spostato nella posizione originaria di *ad*. Troveremo il seguente ciclo-7: *(ad, rd, bd, fd, af, as, ar) (si veda la figura di pagina 147)*.

E gli angoli? Supponiamo di rintracciare il cubetto che era partito da *adf*. Dove l'ha portato *DA?* La risposta è: da nessuna parte; ha fatto un intero giro, ma lungo il cammino si è girato modificandosi in *dfa*. Possiamo denotare questa rotazione oraria - questo quark - con $(adf)_+$. Questo «uniciclo ruotato» è un'abbreviazione del ciclo-3: *(adf, dfa, fad)*. Si può vedere la cosa anche come una rotazione ciclica delle tre lettere *a, d, f* del nome del cubetto. Se il ciclo fosse stato un antiquark, avremmo scritto $(adf)_-$ e le lettere avrebbero ruotato ciclicamente nell'altro senso.

Il cubo è nella posizione di partenza. Le faccette principali dei cubicoli sono indicate dalle *X* nere e i colori principali dei cubetti dalle *O* colorate. (I centri possono essere ignorati in quanto sono stazionari.) Si immagini che le *X* galleggino nello spazio e le *O* siano attaccate al cubo, così che quando le facce girano, le *X* rimangono dove sono mentre le *O* si muovono.

È stata eseguita la rotazione *qF*. I due cerchietti colorati vuoti indicano che quelle faccette hanno perso il loro «equilibrio». Perché esse ritornino all'equilibrio bisognerebbe che una ruotasse di un terzo in senso orario e l'altra di un terzo in senso antiorario. La stessa cosa è avvenuta sulla faccia di sinistra che non si vede.

È stata eseguita la rotazione *qD* dalla posizione di partenza. I cerchietti vuoti ricorrono a coppie. I gialli degli angoli superiore e inferiore hanno rotazioni elidentisi. Gli spigoli blu formano una coppia e lo spigolo giallo è in coppia con uno spigolo rovesciato sulla faccia posteriore non visibile.

Dimostrazione che il rovesciamento è pari e la rotazione è intera.

E gli altri sette angoli? Due di essi - *brs* e *bsf* - rimangono al loro posto e gli altri cinque formano *quasi* un ciclo-5: *(ard, rbd, bfd, saf, ras)*. Purtroppo il ciclo non è del tutto chiuso, perché *ras* arriva sì nel cubicolo originario *ard*, ma girato: arriva in *dar*, che è una rotazione antioraria di *ard*. Questo significa che abbiamo a che fare con un ciclo-15. Questo ciclo è però così vicino al ciclo-5 precedente che ci basterà aggiungere un segno meno per rappresentare la rotazione antioraria. Il nostro ciclo-5 ruotato è allora $(ard, rbd, bfd, saf, ras)_-$ e l'intero effetto di *DA*, espresso in notazione ciclica, è *(ad, rd, bd, fd, af, as, ar)* $(adf)_+$ $(ard, rbd, bfd, saf, ras)_-$.

Ora che abbiamo *DA* in notazione ciclica possiamo effettuare le rotazioni mentalmente, con un semplice calcolo. Quale sarebbe, per esempio, l'effetto di $(DA)^5$? Il cubetto di spigolo *ad* andrebbe avanti di cinque passi lungo il suo ciclo, che lo porterebbe in *as*. (La cosa può essere vista anche come un movimento di due passi all'indietro.) Poi *as* andrebbe in *fb* e così via. Il ciclo-7 è rimpiazzato da un nuovo ciclo-7: *(ad, as, fd, rd, ar, af, bd)*. Prendiamo ora in considerazione il ciclo-5 ruotato. Il cubetto d'angolo *ard* sarebbe portato cinque passi avanti lungo il suo ciclo, tornando su se stesso ruotato in senso antiorario, vale a dire *dar*. Analogamente, tutti i cubetti d'angolo del ciclo-5 tornerebbero alle loro posizioni di partenza, ma con rotazione antioraria; così elevato alla quinta potenza, un ciclo-5 ruotato diviene cinque antiquark. Ma se le cose stanno così, come può essere soddisfatto il requisito della rotazione intera? Non abbiamo un quark - $(ard)_+$ - e cinque antiquark e quindi in totale quattro quark, con una rotazione totale di −1 e 1/3? Bene, vi ho tolto qualcosa. Riuscite a capire cosa? Per impadronirvi bene della notazione ciclica potreste cercare di trovare la rappresentazione ciclica delle varie potenze di *DA* e *AD* e dei loro inversi.

Qualsiasi sequenza di mosse può venire rappresentata in termini di cicli disgiunti di varie lunghezze: cicli senza elementi comuni. Se però volete che i cicli abbiano membri in comune, un ciclo può essere ulteriormente spezzato in cicli-2 (chiamati trasposizioni o a volte scambi). Consideriamo, per esempio, tre animali: un alligatore, una lince e un cammello. Essi occupano inizialmente tre habitat naturali: *A, B* e *C* *(si veda la figura di pagina 148)*. L'effetto del ciclo-3 è quello di mettere i tre animali nel seguente ordine: cammello, alligatore, lince. Lo stesso effetto può però essere raggiunto affettuando prima lo scambio *(A,B)* (ciò che era in *A* va in *B* e viceversa) e poi lo scambio *(A,C)*. Naturalmente, si può raggiungere lo stesso risultato con i due scambi successivi *(A,C)(B,C)* e *(B,C)(A,B)*. Nessuna sequenza di tre scambi, invece, può far rag-

A sinistra il ciclo-4 **(ad, rd, bd, ar)**, *a destra il ciclo-4* **(ad, dr, bd, fd)**.

A sinistra il ciclo-4 **(ad, af, as, ar)**, *a destra il ciclo-4 rovesciato* **(ad, af, as, ar)+**.

giungere lo stesso effetto di *(A,B,C)*. Provate e vedrete. (Si noti che un habitat è come un cubicolo e un animale è come un cubetto.)

Un teorema elementare della zoo-teoria (un campo in cui non ci addentreremo in questa sede) afferma che comunque si riduca una data permutazione di animali tra i loro habitat a un prodotto di scambi successivi (che possono essere sempre fatti) la parità del numero di tali scambi è invariante; vale a dire, una permutazione non può essere espressa come un numero *pari* di scambi una volta e un numero *dispari* un'altra volta. Per di più, la parità di una qualsiasi permutazione è la somma delle parità delle permutazioni in cui è divisa (usando le regole per l'addizione di numeri pari e dispari: dispari più pari è dispari e così via).

Ora, questo teorema ha delle ripercussioni per il Cubo Magico. In particolare ci si può rendere conto che una rotazione *q* consiste di due disgiunti cicli-4 (uno sugli spigoli e uno sugli angoli). Qual è la parità di un ciclo-4? È dispari, come potete accorgervi da soli. Quindi dopo una rotazione *q* sia gli spigoli sia gli angoli hanno avuto un numero dispari di permutazioni, dopo due rotazioni *q* un numero pari, dopo tre rotazioni *q* un numero dispari e così via. Gli spigoli e gli angoli rimangono in fase, nel senso che le parità delle loro permutazioni sono identiche. Ora, chiaramente la permutazione nulla è pari (zero scambi). Così se abbiamo una permutazione nulla sugli angoli, anche la permutazione sugli spigoli deve essere pari. Per converso, una permutazione nulla sugli spigoli implica una permutazione pari sugli angoli. Immaginiamo uno stato identico a quello di partenza tranne per due spigoli che si sono scambiati tra di loro (cioè, uno scambio). Questo stato è pari negli angoli, ma dispari negli spigoli, quindi è impossibile. Al massimo potremmo avere *due coppie* di spigoli scambiati tra loro. La stessa argomentazione vale per gli angoli. In breve, singoli scambi sono impossibili; gli scambi devono avvenire a coppie. (È questa l'origine di uno dei quattro fattori pari a due nel calcolo fatto in precedenza del numero degli stati del cubo.) Ci sono processi per scambiare due coppie di spigoli, due coppie di angoli e anche per scambiare contemporaneamente una coppia di spigoli e una coppia di angoli. (Quest'ultimo processo comporta necessariamente un numero dispari di rotazioni *q*.)

Per completare il discorso sulle limitazioni, esaminiamo le limitazioni su un angolo che gira e su uno spigolo che si rovescia. Una chiara spiegazione ci viene dai matematici John Horton Conway, Elwyn R. Berlekamp e Richard K. Guy, che sviluppano alcune idee di Anne Scott. Il concetto fondamentale è che vogliamo dimostrare che il numero di cubetti rovesciati è sempre pari e che la rotazione è sempre intera. Per stabilire che cosa è rovesciato e che cosa è ruotato, però, abbiamo bisogno di uno schema di riferimento. Per procurarcelo, definiremo due concetti: «faccetta principale di un cubicolo» e «colore principale di un cubetto». (Si ricordi che un cubicolo è una nicchia e un cubetto è un oggetto solido.) La faccetta principale di un cubicolo sarà quella sulla superficie superiore o inferiore del cubo, se c'è una delle due; altrimenti sarà quella sulla parete sinistra o destra *(si veda il cubo in alto della figura a pagina 144)*. Ci sono nove faccette principali su *A*, nove su *B* e quattro sull'equatore. Possiamo invece trascurare i centri perché non sono mai rovesciati o ruotati. Il colore principale di un cubetto è definito come il colore che avrebbe la faccetta principale del cubetto quando il cubetto torna sul suo cubicolo nella posizione di partenza.

Supponiamo ora che il cubo sia mescolato. Un cubetto che abbia il suo colore principale nella faccetta principale del suo attuale cubicolo sarà detto equilibrato; altrimenti sarà detto rovesciato (se è un cubetto di spigolo) o ruotato (se è un cubetto d'angolo). Ovviamente sono due i modi in cui può essere ruotato un cubetto: in senso orario (rotazione +1/3) e in senso antiorario (rotazione −1/3). Il «rovesciamento» di uno stato del cubo sarà definito come il numero di cubetti di spigolo rovesciati in esso contenuti, e la «rotazione» come la somma delle rotazioni degli otto cubetti d'angolo. Diciamo che il rovesciamento e la rotazione dello stato di partenza sono entrambi zero.

Consideriamo ora le 12 possibili rotazioni *q* di cui ogni altro movimento è composto. Effettuando *A* o *B* (o i loro inversi) si mantengono sia il rovesciamento sia la rotazione, dato che nulla lascia la faccia superiore o inferiore e nulla vi entra. *F* o *R* (o i loro inversi) lasciano costante la rotazione, cambiando in una sola volta la rotazione di quattro angoli: due di +1/3 e due di −1/3. Lasciano anche a uno il rovesciamento *(si veda il cubo al centro della figura a pagina 144)*. Anche *S* o *D* lasceranno costante la rotazione (anche questa volta la rotazione di quattro angoli si cancella a coppie) e modificherà di quattro il rovesciamento, dato che quattro cubetti si rovesceranno sempre *(si veda il cubo in basso della figura a pagina 144)*. La conclusione è ciò che abbiamo affermato in precedenza senza dimostrazione: gli otto cubetti d'angolo sono sempre orientati in modo da portare la rotazione totale a un numero intero e i 12 cubetti di spigolo devono essere sempre orientati in modo da rendere pari il rovesciamento totale.

Dopo questa discussione sulle limitazioni, dovreste aver raggiunto la convinzione che, comunque ruotiate e giriate il vostro Cubo Magico, non potete raggiungere, iniziando dalla posizione di partenza, più di un dodicesimo dell'«universo» concepibile. È un'altra questione, invece, dimostrare che ogni stato in quel dodicesimo di universo sia raggiungibile dalla posizione di partenza (ossia, dicendo la stessa cosa all'inverso: che la posizione di partenza sia raggiungibile da ogni stato nel dodicesimo di universo). Per farlo bisogna mostrare come compiere tutte le permutazioni pari di cubetti e come effettuare tutti gli orientamenti che non violino le due limitazioni descritte in precedenza. Ne deriva la necessità di dimostrare che ci sono operatori che compiono sette classi di operazioni: (1) un qualsiasi doppio scambio di coppie di spigoli, (2) un qualsiasi doppio scambio di coppie di angoli, (3) un qualsiasi rovesciamento di due spigoli, (4) un qualsiasi mesone, (5) un qualsiasi ciclo-3 di spigoli, (6) un qualsiasi ciclo-3 di angoli e (7) un qualsiasi barione.

Naturalmente, ciascuno di questi operatori dovrebbe funzionare senza causare effetti collaterali su altre parti del cubo. Con questi potenti strumenti saremmo in grado di coprire senza difficoltà il dodicesimo di universo. Nel caso degli scambi sovrapposti di animali, ho dimostrato che un ciclo-3 consiste in realtà di due cicli-2 sovrapposti. Questo significa che le classi 5 e 6 possono essere ricavate dalle prime quattro classi. Analogamente, un barione può essere ricavato da due mesoni sovrapposti. Abbiamo quindi in realtà bisogno solo delle prime quattro classi.

Per dimostrare che sono disponibili tutti gli operatori appartenenti a queste quattro classi, usiamo un'altra delle idee più importanti e più belle della cubologia: quella relativa agli elementi coniugati. Risulta che tutto ciò che ci serve è un esempio di ciascuna classe; dato un esempio, possiamo costruire tutti gli altri operatori di quella classe. Come funziona la cosa? L'idea è molto semplice.

Supponiamo di aver trovato un operatore della classe 1 che scambi, per esempio, *af* con *ar* e *as* con *ad*, lasciando intoccato il resto del cubo *(si vedano le frecce in colore nella figura di pagina 149)*. Chiamiamolo *H*. Supponiamo ora di voler scambiare due coppie di cubetti di spigolo completamente differenti, per esempio *df* con *bf* e *dr* con *bd* *(si vedano le frecce in nero nella stessa figura)*. Uno pensa: «Se solo quei cubetti fossero nei "quattro magici punti di scambio" sulla superficie in alto...». Bene, perché non metterceli? Dovrebbe essere abbastanza semplice portare quattro cubetti in quattro cubicoli dati. L'ovvia obiezione è: «Sì, ma la cosa avrebbe dei tremendi effetti collaterali - manderebbe completamente all'aria il resto del cubo.» Ma c'è una soluzione intelligente. Chiamiamo *A* l'operazione distruttiva che porta quei quattro cubetti nei punti magici di scambio e supponiamo di avere preso la precauzione di trascrivere la sequenza di mosse di *A*. Subito dopo aver effettuato *A*, effettuiamo il nostro doppio scambio *H*. Ora viene la parte più astuta. Leggendo la nostra trascrizione in ordine inverso e invertendo ogni rotazione *q*, compiamo l'esatto inverso di *A*. Questo non solo riporterà i quattro cubetti nei loro vecchi cubicoli, ma neutralizzerà anche gli effetti collaterali creati da *A*. A questo punto il cubo è ritornato come prima? Non del tutto: dato che tra *A* e *A'* abbiamo effettuato *H*, i quattro cubetti di spigolo tornano in sede permutati tra loro; ciascuno, cioè, finisce nella sede del cubetto con cui è stato scambiato! Tutto il resto è come prima, quindi abbiamo ottenuto proprio quel doppio scambio che intendevamo effettuare.

Se ci si pensa bene, l'operazione appena descritta è concepita in modo impeccabile. La manovra inversa, *A'*, non «sa» che abbiamo scambiato due coppie di spigoli: per quanto la riguarda, sta semplicemente riportando tutto al punto in cui era prima di *A*. Abbiamo quindi fatto il nostro scambio sotto il naso di *A'*; come dire che abbiamo «imbrogliato il cubo». In termini simbolici, abbiamo compiuto la sequenza di mosse *AHA'*, sequenza che viene detta un «coniugato» di *H*.

È questo tipo di illustrazione concreta di un'astratta nozione della teoria dei gruppi che fa del Cubo Magico una delle cose più affascinanti che siano mai state inventate per l'insegnamento della matematica. Normalmente, nella teoria dei gruppi, gli esempi di elementi coniugati sono o troppo banali o troppo astratti per illuminare o interessare. Il Cubo Magico fornisce una vivida illustrazione di elementi coniugati e di molti altri importanti concetti della teoria dei gruppi.

Supponiamo che vogliate sistemare una coppia quark-antiquark su angoli opposti, ma sappiate farlo solo su angoli adiacenti. Come procedere? Ecco un suggerimento: ci sono due belle soluzioni, ma la più breve ed elegante comporta l'uso di un coniugato. Per inciso, una manovra che crei un quark su un angolo (insieme, naturalmente, ad altri effetti) potrebbe essere chiamata un quarkgiro.

Quello che abbiamo mostrato per gli spigoli funziona anche con gli angoli: se si riesce a scambiare due angoli *particolari* si è in grado anche di scambiare due angoli *qualsiasi*. La coniugazione ci permette di costruire un'intera classe di operatori a partire da un singolo membro di quella classe. Naturalmente rimane la questione: come trovare degli esempi di operatore per ciascuna delle quattro classi? Per esempio, come trovare un operatore che crei un mesone su due angoli adiacenti (una combinazione di un quarkgiro e di un antiquarkgiro)? Come trovare un operatore che scambi due coppie di spigoli che si trovino entrambe sulla superficie in alto? Non darò qui la risposta ma, seguendo Singmaster, suggerirò l'esplorazione quasi sistematica di alcuni piccoli «sottouniversi» all'interno della totalità degli stati del cubo; il che equivale a suggerire di prendere in considerazione dei *sottogruppi*. Questo significa limitare deliberatamente il vostro insieme di mosse ad alcuni particolare tipi di mossa. Dò qui di seguito cinque esempi di interessanti sottogruppi creati da vari tipi di restrizione:

1. Il gruppo fetta. In questo sottogruppo, ogni rotazione di una faccia deve essere accompagnata dalla mossa parallela sulla faccia opposta. Così *D* deve essere accompagnato da *S'*, *A* da *B'* e *F* da *R'*. Il nome viene dal fatto che una mossa doppia di questo tipo equivale alla rotazione di una delle tre fette centrali del cubo. Singmaster abbrevia la mossa fetta *DS'* con D_s, *D'S* con D_s' e così via. Con questa restrizione le facce non possono essere mescolate in modo arbitrario: ogni faccia avrà una disposizione in cui tutti e quattro gli angoli hanno lo stesso colore. Un caso speciale è la disposizione chiamata Punti, in cui ogni faccia è di un unico colore tranne che per il centro *(si veda la figura a pagina 139)*. Riuscite a capire come si può raggiungere la posizione Punti dalla posizione di partenza? Quanti differenti modi ci sono per disporre i punti? In cosa assomiglia la posizione Punti a un mesone? (Nota: il lettore troverà le risposte a tutte queste domande, insieme a molte altre, nel libro di Singmaster. Per favore non indirizzate a me le risposte; mandate a Singmaster tutte le intuizioni innovative che vi possono venire.)

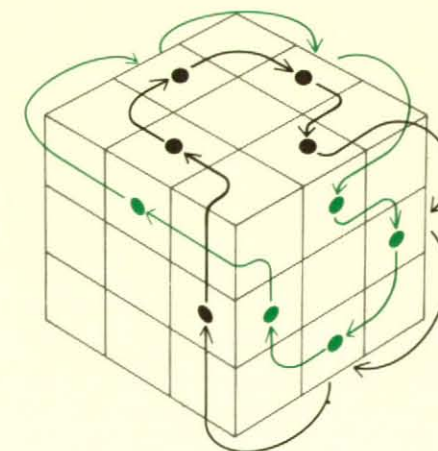
2. Il gruppo fetta quadrato. È questa un'ulteriore restrizione del gruppo fetta, in quanto sono consentiti solo quadrati di mosse fetta, quali D_s^2 (che equivale a D^2S^2) o F_s^2 (che equivale a F^2R^2).

3. Il gruppo antifetta. Qui, invece di ruotare sempre in parallelo le facce opposte, le ruotiamo sempre in antiparallelo, così che *D* è accompagnato da *S*, *F* da *R* e *A* da *B*. Una mossa antifetta ha una *a* sottoscritta: D_a equivale a *DS*. (Naturalmente, il gruppo antifetta quadrato non differisce dal gruppo fetta quadrato.)

4. Il gruppo due facce. Vi permette di ruotare solo due facce, ad esempio *F* e *D*.

5. Il gruppo due quadrati. Come nel gruppo due facce, si possono ruotare solo due facce, ma per di più si possono compiere solo rotazioni di 180 gradi. Si tratta di un sottogruppo molto semplice.

Prendendo in considerazione i gruppi due facce e due quadrati, si riuscirà a trovare processi che effettuano doppi scambi, alcuni di spigoli e altri di angoli. È interessante notare che questi processi da soli, insieme alla nozione di coniugazione, ci consentono - in linea teorica - di risolvere interamente il problema del ripristino dell'ordine.

Perché non abbiamo bisogno anche di

Il ciclo-7 **(ad, rd, bd, fd, af, as, ar)**.

un produttore di mesoni e di un doppio rovesciatore di spigoli? Bene, consideriamo come si potrebbe arrivare a un doppio rovesciatore di spigoli partendo dalle due classi di strumenti che possiamo presupporre trovate. Per rovesciare due spigoli senza creare effetti collaterali, effettueremo due successivi doppi scambi di coppie di spigoli, entrambi coinvolgenti le stesse coppie! Per esempio, potremmo scambiare af con ar e bf con br e poi scambiarli nuovamente. Questo sembra un «processo nullo», ma non è detto che sia così. Dopo tutto, come prima, possiamo inserire il secondo scambio tra un processo X e il suo inverso X', scegliendo accuratamente il processo X in modo che... (Oh cielo! non riesco più assolutamente a seguire il corso dei miei pensieri. Sono sicuro, però, che voi riuscirete a completarlo. Ricordo che non si trattava di una cosa complicata e che l'idea mi sembrava avere una certa eleganza. Sono certo che sembrerà così anche a voi.)

Con lo stesso tipo di ragionamento si può arrivare al modo per costruire un produttore di mesoni partendo da semplici processi di scambio d'angoli e dalla coniugazione. Dati i mesoni si possono costruire i barioni. E se si hanno a disposizione mesoni, barioni, doppi rovesciatori di spigoli, doppi scambiatori di coppie di spigoli e doppi scambiatori di coppie d'angoli, si ha un intero set di strumenti con cui riportare qualsiasi cubo mescolato alla posizione di partenza, sempre che appartenga alla stessa orbita della posizione di partenza. È questa, non c'è bisogno di dirlo, una prova d'esistenza altamente teorica e l'organizzazione pratica di un insieme di routine sarebbe molto diversa. Il tipo di soluzione descritto ha il pregio di avere una descrizione compatta, ma ha un'efficacia minima. In pratica, un solutore deve costruirsi un insieme molto ampio e versatile di routine che siano brevi, facili da memorizzare e ampiamente ridondanti. È chiaramente vantaggioso essere in grado di operare le trasformazioni richieste in più modi: si può scegliere lo strumento che sembra più adatto per la situazione concreta, invece di dover usare, per esempio, uno strumento teorico che arriva a costruire un barione in parecchie centinaia di rotazioni q.

Il solutore tipo elabora un insieme di processi in parte per intuizione, in parte per caso, a volte con l'aiuto di diagrammi e all'occasione con principi astratti di teoria dei gruppi. Un principio che quasi tutti formulano abbastanza presto è quello di «mettere le cose fuori posto». Si tratta sempre dell'idea dei coniugati, solo in forma semplificata. In genere l'operazione è accompagnata da osservazioni del tipo: «Vediamo, metto fuori posto *questo* così posso muovere *quello*, e ora posso girare di nuovo *questo*. Là, ora *quello* è dove volevo». Ci si sente chiaramente la struttura coniugata.

L'unico problema dell'essere coscienti del modo in cui funziona ciò che si sta facendo è che può essere troppo faticoso. Io ho l'impressione che la maggior parte dei cubologhi non pensi in dettaglio al modo in cui loro strumenti raggiungono il loro obiettivo, almeno non mentre sono impegnati a ripristinare l'ordine di qualche cubo mescolato. I solutori esperti sono piuttosto come dei virtuosi del pianoforte che hanno memorizzato dei pezzi difficili. Come mi diceva un cubologo del MIT: «*Io* ho dimenticato come risolvere il problema, ma le mie *dita* no».

Io, probabilmente, procedo troppo alla cieca. Non solo non mi chiedo perché funzionino i miei operatori mentre li uso; per alcuni di essi devo ammettere che non ho proprio la più pallida idea del perché funzionino. Ho trovato questi «magici operatori» con un lungo e penoso procedimento per tentativi ed errori e ho usato degli approcci euristici: «Tentiamo con diverse potenze di sequenze semplici», «Usiamo un po' di coniugati», e così via. Ciò di cui non mi sono quasi mai servito - mi spiace per il povero Rubik - è stata la visualizzazione tridimensionale. Conosco però un cubologo di Standford, Jim McDonald, che può giustificare ogni rotazione q che compie. I suoi operatori non gli sembrano magici perché riesce a capire quello che stanno facendo in ogni momento. In realtà, non li ha memorizzati come ho fatto io per i miei; sembra che li ricostruisca mentre lavora alla soluzione, confidando sul suo «senso del cubo». È come un musicista esperto che può improvvisare là dove un principiante deve ricordare. Se a qualche lettore può interessare, l'idea centrale di Jim consiste nel risolvere per prima cosa lo strato superiore, meno un angolo, e poi tenere una colonna verticale da usarsi allo stesso modo in cui si usa un vialetto laterale per girare la macchina. Gli altri due strati vengono risolti smistando i cubetti dentro e fuori il «vialetto».

Forse non è un caso che l'approccio astratto sia stato portato al suo estremo da un collega di Singmaster, Morwen B. Thistlethwaite, che detiene attualmente il record mondiale per il più corto algoritmo in grado di risolvere il problema del ripristino dell'ordine. Tale algoritmo richiede al massimo 52 «rotazioni». (Una rotazione può essere o una rotazione q o una mezza rotazione, cioè una rotazione di 180 gradi di una faccia.) Thistlethwaite ha utilizzato alcuni concetti della teoria dei gruppi per guidare un calcolatore alla ricerca di particolari tipi di trasformazioni. Il suo algoritmo ha la proprietà di non sembrare affatto convergere, fino alle ultimissime rotazioni, verso lo stadio finale della soluzione.

***Un alligatore, una lince e un cammello* (a, b, c) *si sono scambiati gli habitat naturali* (A, B, C).**

Illustriamo ora, per contrasto, uno stile più convenzionale. La maggior parte degli algoritmi iniziano col far risultare interamente corretto uno strato, per esempio quello superiore. (Parlando di «strato superiore», invece che di «superficie superiore», intendo che anche l'«orlo» deve essere giusto; i cubetti in alto devono essere giusti sia visti di fianco sia visti dall'alto.) Questo rappresenta il primo di una serie di «stati base». Anche se successivamente uno stato base deve essere distrutto, quello stato sarà più avanti ripristinato, e ogni volta che questo succede si sarà introdotto un ordine maggiore. Sono questi i successivi stati base.

Dopo aver finito con lo strato superiore, in genere il solutore lavora sugli angoli dello strato inferiore, oppure cerca di sistemare l'equatore orizzontale. In realtà, la maggior parte degli algoritmi può essere suddivisa in cinque fasi naturali, corrispondenti alle classi naturali dei cubetti che tornano nei loro cubicoli di base. Il mio algoritmo personale, per esempio, passa attraverso le cinque seguenti fasi: spigoli superiori, angoli superiori, angoli inferiori, spigoli dell'equatore e spigoli inferiori. Nelle prime due fasi la posizione e l'orientamento sono raggiunti simultaneamente. Le ultime tre fasi si suddividono ciascuna in una fase di posizionamento e una di orientamento. Naturalmente, gli operatori di una fase devono rispettare tutto ciò che è stato realizzato nelle fasi precedenti. Questo significa che possono danneggiare l'ordine costruito solo se poi vi pongono rimedio; nulla vieta, invece, che mettano per aria a piacimento dei cubetti che andranno messi in gioco in fasi successive. Mi sembra che gli algoritmi utilizzati dagli altri solutori siano in genere basati sulle stesse classi di cubetti, ma l'ordine in cui si succedono le varie fasi può essere molto diverso.

Virtualmente, tutti gli algoritmi hanno la proprietà che se si scattassero una serie di fotografie a un cubo nei suoi stati base, si vedrebbero interi gruppi di cubetti andare al loro posto secondo schemi. Questo fatto prende il nome di monotonicità a livello degli operatori: un visibile e costante avvicinamento alla posizione di partenza, senza arretramenti. Naturalmente, si ha arretramento a livello delle *rotazioni*, ma questo è un altro discorso.

Di spirito molto differente è l'algoritmo di Thistlethwaite. Invece di cercare di mettere determinate classi di cubetti nei loro cubicoli, egli compie una «discesa per sottogruppi sempre più piccoli». Questo significa che parte effettuando qualche mossa in piena libertà, poi riduce i tipi di mosse consentite d'ora in avanti, compie ancora qualche mossa, riduce ancora un po' e così via finché le limitazioni sono tali da non consentire più alcuna mossa. A quel punto, però, si è raggiunta la posizione di partenza! Ogni volta, la riduzione equivale a proibire rotazioni q su due facce opposte, consentendo solo, al loro posto, mezze rotazioni. Le prime facce a essere così «imbrigliate» sono A e B, poi vengono F e R e infine S e D. La cosa strana di questo metodo è che non ci si accorge di avvicinarsi alla posizione di partenza, neanche scattando una serie di foto in momenti accuratamente scelti: tutto succede all'improvviso.

Viene naturale chiedersi a questo punto: esiste un semplice modo per stabilire quanto si è lontani dalla posizione di partenza? La cosa sarebbe ovviamente molto utile. È piuttosto imbarazzante, per esempio, dover ricorrere a tutta la forza di un algoritmo generale per annullare quello che un amico ha fatto con quattro o cinque rotazioni a caso. Non fosse che per questo, sarebbe bello riuscire a stabilire rapidamente se un certo stato è «proprio casuale» o è vicino alla posizione di partenza. Ma cosa significa «vicino»? In questo vasto spazio, la distanza tra due stati può essere misurata in due modi abbastanza naturali: si può contare il numero di rotazioni q oppure il numero di rotazioni necessarie per passare da uno stato all'altro (dove «rotazione», come prima, sta per rotazione q o per mezza rotazione). Ma come si può sapere quante rotazioni sono necessarie per raggiungere la posizione di partenza senza compiere una ricerca esaustiva? Sarebbe preferibile una semplice stima attendibile e abbastanza accurata, una stima realizzabile rapidamente con una veloce ispezione dello stato in cui si trova il cubo. Si potrebbe ingenuamente suggerire di contare il numero di cubetti che non si trovano nei loro cubicoli base; ma per capire che la cosa non funziona affatto basta pensare alla posizione Punti, in cui quasi tutti i cubetti sono sul lato «sbagliato»: quella posizione è a solo otto rotazioni q dalla posizione di partenza. Forse per una stima migliore si potrebbero prendere in considerazione il rovesciamento e il numero di quark, ma non so se la cosa sia possibile.

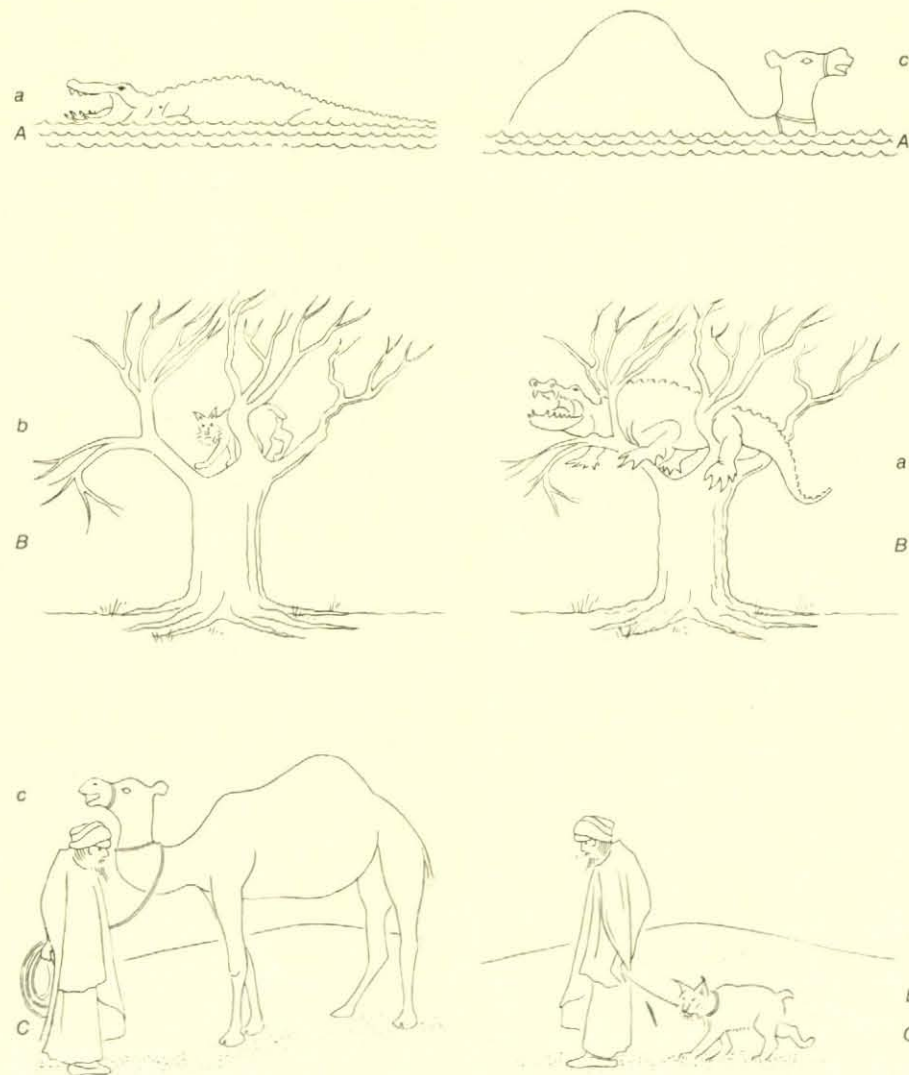
Sofisticate argomentazioni basate sulla teoria dei gruppi stimano che non si possa essere lontani dalla posizione di partenza più di 22 o 23 rotazioni. La cosa è decisamente sorprendente se si pensa che i primi algoritmi della maggior parte dei solutori comportano parecchie centinaia di rotazioni e che algoritmi altamente elaborati ne comportano qualcosa come 80-90. In effetti, molti semplici operatori richiedono un numero di rotazioni notevolmente superiore a quello dell'intero algoritmo di Thistlethwaite. (Il mio primo doppio rovesciatore di spigoli era lungo quasi 60 rotazioni.)

Una cosa che sappiamo, e che può essere facilmente dimostrata, è che esistono stati lontani almeno 17 rotazioni dalla posizione di partenza. L'argomentazione segue la seguente linea. All'inizio ci sono 18 possibili rotazioni: S, S', S^2, D, D', D^2 e così via. Dopo di ciò ci sono 15 rotazioni ragionevoli da fare. (Per non ruotare di nuovo la stessa faccia.) Il numero di distinte rotazioni di lunghezza 2 è perciò 18×15, ossia 270. Un'altra rotazione aggiungerà un altro fattore pari a 15 e così via. Quanto ci vuole per arrivare al numero di stati raggiungibili? Risulta che 17 rotazioni creano più sequenze di rotazioni di quanti siano gli stati distinti e che 16 sono troppo poche. Ora, non tutte le sequenze di rotazioni portano a uno stato distinto, e così non abbiamo dimostrato che con 17 rotazioni si arriva a ogni stato raggiungibile. Abbiamo semplicemente dimostrato che ci vogliono *almeno* 17 rotazioni per raggiungere ogni stato dalla posizione di partenza. È quindi concepibile che due stati qualsiasi non siano distanti uno dall'altro più di 17 rotazioni. Ma quali sono queste 17 rotazioni? Questo è il problema.

Come scambiare coppie qualsiasi di spigoli.

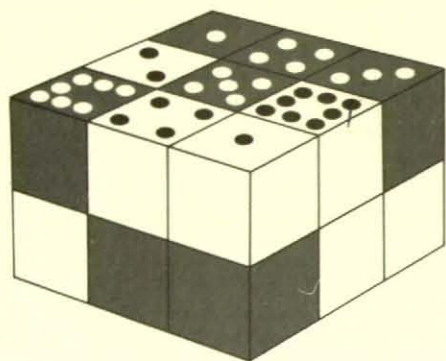
Per ora, Dio solo sa come passare da uno stato all'altro del cubo nel minor numero possibile di mosse. Potremmo allora chiamare «divino» un algoritmo che ci consentisse di saperlo. Scottante problema di cubologia: l'algoritmo divino è una tavola gigantesca senza alcuna regolarità o al suo interno c'è una sufficiente regolarità perché da esso si possa ricavare un breve ed elegante algoritmo padroneggiabile da un comune mortale?

Se Dio partecipasse a una gara per la soluzione del Cubo Magico, potrebbe incontrare un'accanita resistenza da parte di alcuni prodigiosi mortali, che pure non conoscono il suo algoritmo. Mi hanno raccontato che c'è un giovane inglese di Nottingham, di nome Nicholas Hammond, che ha abbassato il suo tempo medio di soluzione a qualcosa come 30 sencondi! Una performance così fenomenale richiede parecchi tipi di abilità. La prima è una profonda conoscenza del cubo. La seconda è un insieme estremamente raffinato di operatori. La terza è una padronanza così completa degli operatori da poterli applicare dormendo. La quarta è la capacità di compiere ad altissima velocità tortuosi movimenti con le mani. La quinta è di avere un ben oliato «cubo da corsa»: uno che giri al minimo tocco delle dita, anticipando quasi ogni operatore. In breve, il cubo da corsa è un cubo che *vuole* vincere.

Non ho mai sentito di alcuno che abbia dato un nome ai suoi cubi da corsa, ma sicuramente accadrà. Si direbbe, comunque, che ci sia un'associazione tra nomi pittoreschi e coloro che hanno dato i maggiori contributi alla cubologia. Oltre

a Singmaster e Thistlethwaite c'è la signora Kathleen Ollerenshaw (recentemente eletta sindaco di Manchester), la quale ha scoperto un gran numero di raffinati processi, ha scritto un articolo sul Cubo Magico e si è distinta per essere stata la prima a soffrire di un grave attacco di dito del cubista, una grave forma del disturbo di cui si è parlato all'inizio dell'articolo. C'è poi Oliver Pretzel, scopritore di un delizioso ciclo-3 ruotato e creatore di un'incantevole «bella disposizione» chiamata stadio 6-*U*, che può essere raggiunta dalla posizione di partenza con la parola $S'D^2F'S'R'ARSFDA'DSD_sF_sA_sD_s$.

Le belle disposizioni sono interessanti per molti appassionati del cubo, ma in questa sede non posso occuparmene a lungo. Citerò solo alcune delle migliori. La mia favorita è l'Animata, il cui «genotipo», o sequenza di rotazioni, è $DAF^2B'D_sF_sB'F'D'F^2DA^2FD^2F'D'A'F'A^2FD$. C'è poi il Drago, una disposizione sinuosa che si snoda attorno al cubo: $RD_sB'D^2BD_sR'D^2AR^2A'BD^2B'$. Se si taglia via la coda del Drago D^2B' e si attacca invece $R^2D_aA^2D_a'R^2B'$, si crea una curiosa disposizione a due anelli. Tutte le precedenti disposizioni sono state inventate dal cubologo Richard Walker.

Un'interessante disposizione è il Mesone Gigante, fatto di un quark gigante (un sottocubo d'angolo 2×2×2 ruotato di 120 gradi) e un antiquark gigante. Per dargli il tocco finale si possono usare quarkgiri per ruotare un quark e un antiquark di dimensioni standard sugli angoli del quark e antiquark gigante, come ciliegine sul gelato *(si veda la figura di pagina 139)*. Lascio a voi il compito di trovare il modo per costruirlo. Un buon esercizio di riscaldamento è scoprire il modo per arrivare allo stadio Pons Asinorum (Ponte degli asini) *(si veda la stessa figura)*, così chiamato perché, come mi ha fatto notare un cubologo del MIT, «se uno non riesce a risolvere questo, può lasciar perdere il cubo». Ci sono poi due tipi di croci, note alla comunità di cubologhi del MIT come la croce di Natale e la croce di Plummer *(si veda la stessa figura)*. La croce di Natale mette in gioco tre coppie di colori (*A-B*, *F-D* e *S-R* nella figura); la croce di Plummer ne mette in gioco due triple nello stile quark-antiquark.

Vorrei lasciare il lettore con qualche suggerimento e qualche proposta di riflessione. Un difficile esercizio, buono per cubisti di qualsiasi livello, è far dare a qualcuno un piccolo numero di rotazioni a un cubo primitivo, poi cercare di tornare da questo stadio moderatamente mescolato alla posizione di partenza trovando l'esatto processo inverso. Gli esperti riusciranno a invertire un numero di rotazioni maggiore dei principianti. Kate Fried riesce a invertire normalmente sette rotazioni, e una volta, dopo aver passato un'intera giornata a osservare il cubo, riuscì a invertirne 10. (Io arrivo fino a circa quattro rotazioni.)

La mia via maestra per scoprire un algoritmo si basa su due stimolanti esercizi che coinvolgono solo cubetti d'angolo. L'esercizio preliminare è il seguente. Portare i quattro cubetti d'angolo bianchi sulla faccia superiore con le faccette bianche rivolte verso l'alto, senza preoccuparsi di quale cubicolo vada a occupare ciascun cubetto. Contemporaneamente fare la stessa cosa sulla faccia inferiore (naturalmente con il *suo* colore rivolto verso il basso). Più complesso è l'esercizio che, oltre a richiedere le stesse cose del precedente, richiede anche di assicurarsi che tutti i cubetti d'angolo siano nei loro giusti cubicoli. Ciò equivale a risolvere il Cubo Magico 2×2×2 e vi farà avanzare notevolmente verso la completa padronanza del Cubo Magico.

Per aiutarvi nei vostri processi sugli spigoli, accennerò, lasciando a voi il compito di approfondirlo, a un meraviglioso trucco scoperto da David Seal e basato su un tipo di operatore chiamato monoflip. Come si può ottenere un doppio rovesciatore di spigoli da un processo che rimescola i due strati inferiori, ma lascia invariato lo strato superiore tranne che per il rovesciamento di un cubetto di spigolo? Vi darò un suggerimento: la risposta mette in gioco l'importante concetto, ricavato dalla teoria dei gruppi, di commutatore, una parola della forma $PQP'Q'$. Lascio a voi anche il compito di trovare il vostro personale operatore monoflip. Personalmente, una volta scopertolo ho incorporato il trucco nel mio metodo.

Ecco un piccolo indovinello: perché i cicli 5 e 7 portano così spesso a un oggetto con simmetrie che hanno tutte a che fare con numeri quali 3, 4, 6 e 8? Da dove vengono fuori lunghezze cicliche come 5 e 7? In qualche modo collegata alla precedente è la domanda seguente: qual è il massimo ordine che una parola può avere? L'ordine di una parola è la potenza a cui la si deve elevare per ottenere l'identità. (Per esempio, l'ordine di *D* è 4.) Si può dimostrare che l'ordine di *DA*, ad esempio, è 105 osservandone la struttura ciclica.

Cosa ci riserverà il futuro della cubologia, di cui in questo articolo ho appena scalfito la superficie? Rubik e altri stanno lavorando su vari tipi di generalizzazioni. Esiste già un Domino Magico, che corrisponde a due terzi di un cubo magico: due strati 3×3. Lo si può muovere solo con mezze rotazioni su due dei suoi tre assi e con rotazioni *q* sul terzo asse. Nella sua posizione di partenza, una faccia è nera, l'altra è bianca ed entrambe le facce sono numerate in ordine da 1 a 9. Assomiglia ancor più del cubo al 15 Puzzle. Molte persone hanno costruito cubi 2×2×2 e forse un giorno cubi di questo tipo saranno messi in vendita. Volendo, potete farvene uno incollando dei piccoli cappucci a tre angoli su ciascuno degli otto angoli di un cubo 3×3×3. I lettori si chiederanno naturalmente quali allettanti possibilità abbia un cubo 4×4×4; rassicuratevi, ne stanno progettando la costruzione in Olanda e sarà forse pronto tra breve. È inevitabile che ci si interroghi sulle dimensioni superiori e inferiori; i teorici del cubo stanno iniziando a discutere le proprietà dei cubi di dimensioni maggiori.

Le potenzialità del cubo 3×3×3 sono lungi dall'essere esaurite. Una ricca area di terreno inesplorato è quella della colorazione alternata, di cui mi hanno parlato vari cubologhi del MIT. Si possono colorare i cubetti in un gran numero di modi diversi e ogni tipo di colorazione presenta un nuovo tipo di problema del ripristino dell'ordine. In una variante, l'orientamento dei cubetti di spigolo diventa irrilevante e assume vitale importanza quello dei cubetti di centro. In un'altra variante diventa irrilevante l'orientamento dei cubetti d'angolo e importante quella dei centri. Tendendo alla semplificazione, si possono colorare due facce nello stesso modo, riducendo così di uno il numero di colori distinti. Oppure si possono colorare le facce con appena tre colori. Al limite si potrebbero avere tre facce blu che si incontrano in un angolo e tre facce bianche che si incontrano nell'angolo diagonalmente opposto. Il funzionario del governo francese di cui ho parlato in precedenza sostiene che i cubi da lui visti avevano cinque facce di un unico colore e la sesta di un altro colore!

Chi può sapere a quali limiti si arriverà? Come ha rilevato Bernie Greensberg: «Il cubismo richiede al potenziale cubista di inventare letteralmente una scienza. Ogni solutore o solutrice deve suggerirsi aree di ricerca, progettare esperimenti, trovare principi, costruire teorie, rigettarle e così via. È l'unico rompicapo che richieda al suo solutore di costruire un'intera scienza.» Rubik e Ishige avrebbero mai potuto immaginarsi che la loro invenzione avrebbe portato a un modello e a una metafora per tutto ciò che di profondo e di bello vi è nella scienza? È veramente sorprendente questo Cubo Magico.

Vorrei ringraziare i molti cubologhi e cubisti che hanno direttamente o indirettamente contribuito ad accrescere le mie conoscenze e a stimolare il mio entusiasmo. Tra essi vi sono Clark Baker, Alan Bawden, Jim Boyce, Larry Breed, Charles Brenner, Bob Filman, Carl Hoffman, Scott Kim, Bill McKeeman, Jeannine Moseley, Richard Pavelle, Dave Plummer, David Policansky, David Singmaster, Ann Trail, Allen Wechsler, Dan Weise e John Woodcock. Ma i miei più caldi ringraziamenti sono riservati a Greenberg, che ha organizzato per me al MIT un incontro di cubologhi e che mi ha contagiato con il suo entusiasmo.

Il Domino Magico in uno stato mescolato.